MATDAT18: Materials and Data Science Hackathon

Team Composition (2 people max.)

Name	Department	Institution	Email
Benjamin Afflerbach	Materials Science & Engineering	University of Wisconsin – Madison	bafflerbach@wisc.edu

Project Title

Dilute Solute Diffusion

Project Synopsis (approx. 100 words)

Solute diffusion, which is the way impurities are transported in alloys, is a critical factor in determining many materials properties such as precipitation, high-temperature creep, and phase transformations. Using a previously generated computational database of solute diffusion characteristics for FCC, BCC, and HCP materials we have generated multiple machine learning (ML) models for predicting solute diffusion activation energies in new materials systems. We are now looking for ways to improve our models by including more advanced machine learning and data science techniques.

Identified Data-Science Collaborative Need (approx. 100 words)

There are a few potential collaborations we would like to focus on. First, in previously generated ML models we have analyzed overall performance through cross validation RMSE. We would like to obtain error estimates that are better predictors of new systems, e.g., materials compositions that are not well sampled within the dataset. Second, we would like to improve our methods for descriptor generation and selection. We currently generate elemental descriptors using combinations of elemental properties (e.g., melting temperature, atomic radius) and select them using various Scikit-learn feature selection routines such as recursive feature elimination and univariate feature selection. We would like to explore alternate ways to reduce the number of features, combine elemental descriptors automatically, and overall understand which descriptors are the most important. Finally, we are interested in design of experiments approaches to guide future calculations to obtain the lowest possible errors on different target systems.

Data Origin and Access (*data must be available and sharable with data science teams* – please address: data source/origin, access privileges, sharing privileges)

The dataset was generated in the Computational Materials Group at UW-Madison and is published in Wu et al., Scientific Data 3, 160054 (2016). The dataset is freely available under a CC BY 4.0 license on figshare, and available for perusal at the CMG group webpage.

https://figshare.com/articles/DFT_dilute_solute_diffusion_in_Al_Cu_Ni_Pd_Pt_and_Mg/1546772

https://matmodapp.engr.wisc.edu/https-only/dsd_calculated/cmg_dilute_solute_diffusion_main.php

Project Description (approx. 1.5 pages, plus figures and references; please describe data size, form, dimensionality, uncertainties, number of examples, etc.)

The dataset we wish to use contains 375 host-impurity diffusion activation energy barriers calculated by DFT methods. The energy barriers describe the movement of the impurity element within the host material. Each diffusion activation energy is normalized by subtracting off the host self-diffusivity. Therefore, the key prediction obtained from the machine learning model is the diffusion activation energy relative to the host's value. There are 3 crystal systems across 14 host materials. 10 of the host materials belong to the FCC crystal system, one is an HCP structure, and three hosts are BCC structures. The dataset initially consists of 3 features, the host material, the solute material, and the activation energy barrier. Using known databases of elemental properties, we can generate ~600 potential descriptors to use in our machine learning efforts. These descriptors fall into six main categories. These are: Host property, Solute property, Average property, Difference in property, Minimum property, and Maximum property. The host and solute properties are simply the elemental property of the host and solute element respectively. The other four descriptors are logical or mathematical combinations of the host and solute descriptors. This forms the complete dataset that we hope to explore. Previous machine learning efforts have been done on a subset of this dataset that only includes the FCC structures and the results are published in H. Wu, A. Lorenson, B. Anderson, L. Witteman, H. Wu, B. Meredig, D. Morgan Robust FCC solute diffusion predictions from ab-initio machine learning methods Computational Materials Science, 134, 160-165 (2017)