
Ryan Jacobs, Tam Mayeshiba, Ben Afflerbach, Dane Morgan
(University of Wisconsin – Madison, WI USA)

Luke Miles, Max Williams, Matthew Turner, Raphael Finkel
(University of Kentucky, Lexington, KY USA)

Most Recent Skunkworks MASTML members:
Avery Chan, Hock Lye Lee, Min Yi Lin

https://github.com/uw-cmg/MAST-ML

NanoHub ML Workshop
5/19/2021
Machine learning in Materials Science is Exploding

A Basic Materials Design Workflow

1. Identify Materials Properties
2. Train Model of Properties
3. Predict Properties for New Chemical Compositions
4. Synthesize and Verify Predictions

Training Details

1. Generate Training Data
2. Data Cleaning
3. Feature Generation and Engineering
4. Model Assessment
5. Model Optimization
6. Predictions
What is MAST-ML?

MAST-ML is an open-source Python package designed to broaden and accelerate the use of machine learning in materials science research, particularly for non-experts.

https://github.com/uw-cmg/MAST-ML
MAST-ML automates the supervised learning workflow

- MAST-ML supports the full library of scikit-learn modules, and can be used to construct neural networks with Keras (based on tensorflow)

- MAST-ML allows for the simultaneous execution of an arbitrary combination of data preprocessing, feature generation/selection, model types and model evaluation metrics
(NSF CSSI) Machine Learning Materials Innovation Infrastructure

(PIs Dane Morgan, Paul Voyles, Michael Ferris, Ryan Jacobs, Ben Blaiszik)
(NSF CSSI) Machine Learning Materials Innovation Infrastructure

Data for Model

Organize Data
- User-Provided Data

Open Materials Databases
- AFLOW
- Materials Project

Globus-Accessible Data
- Khazana
- imatN

Model Building and Evaluation

Foundry Computing
- MAST-ML
- Keras, Tensorflow, SciKitLearn, PyTorch

Data access: MAST Forge

Cloud computing: Jetstream, AWS

Model Hosting and Sharing

Disseminate Results
- Model Publication

DLHub
- Data and learning hub for science

Local Model Analysis
- Data Publication

Code Publication
- Github

MAST-ML

Model building, evaluation, and key connections between data and model dissemination
Diffusion of dilute impurity X in host H. We have DFT calculations of 440 values, but want ~4,000. [1, 2]

Assume Y= Activation energies measured relative to host, X= Host descriptors, Impurity descriptors. Find Y=F(X).

Descriptors = elemental properties like melting temperature, bulk modulus, electronegativity, ... and their ratios, differences, etc. (MAGPIE set)[3]

F is determined using standard machine learning regression methods (e.g., Gaussian Process Regression (Gaussian Kernel) (GPR), Random Forest (RF), neural network).

Fit F with calculated data (15 hosts, 440 M-X pairs)

Getting Started with the MAST-ML tutorial on NanoHub

• Link to Tool: https://nanohub.org/tools/mastmltutorial
  • Select “Launch Tool”
  • A Jupyter notebook environment will open (may take a minute)
  • Click on cell and run with Shift+return
  • Data will be saved to local directory, see next slides for how to download results