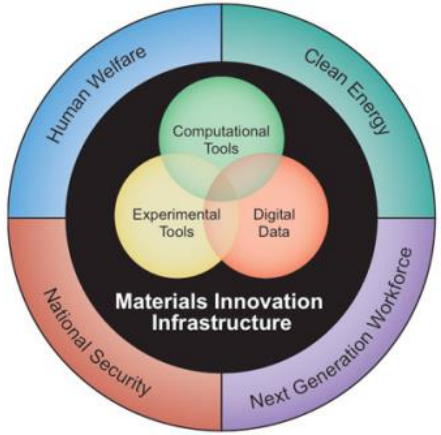
NIST
National Institute of Standards and Technology
U.S. Department of Commerce

JARVIS-ML

2D/3D materials screening and genetic algorithm with ML model

Kamal Choudhary, Brian DeCost, Francesca Tavazza
University of Maryland

August 02, 2018



JARVIS

Acknowledgement and Collaboration

- Carelyn Campbell, Daniel Wheeler, Aaron Gilad Kusne, Jason Hattrick-Simpers, Martin Green, Faical Y. Congo, Zachary Trautt, Andrew Reid, Nhan Van Nguyen, Sugata Chowdhury, Albert Davydov, Irina Kalish, Chandler Becker, Ryan Beams, Adam Biacchi, Kevin Garrity, Russell Johnson, John Vinson, NIST
- Logan Ward, University of Chicago
- Ankit Agrawal, Northwestern University
- Patrick Riley, Google
- Yuri Mishin, George Mason University
- Richard Hennig, University of Florida
- Tao Liang, Penn state
- Materials-Project team, Lawrence Berkeley National Laboratory
- Evan Reed, Stanford University
- Xin Zhao, Fen Zhang, Ames lab
- Sam Reeve (nanohub.org), Purdue University
- Lidia Carvalho Gomes, National University of Singapore

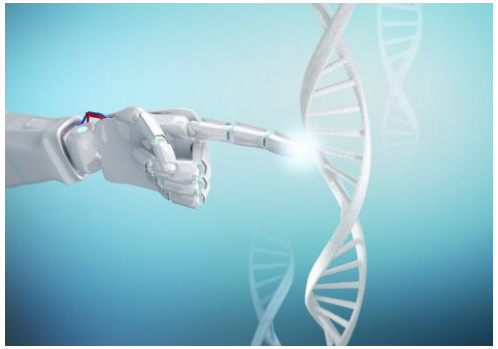
Outline

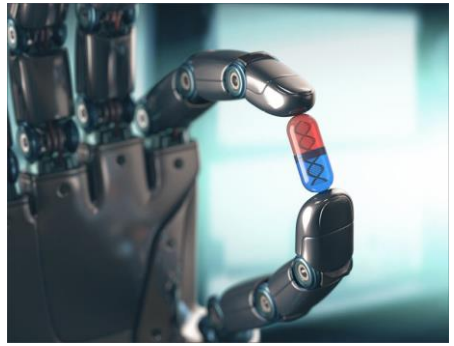
- AI in materials, Motivation
- JARVIS-FF, DFT and ML
- Representing materials to computers
- Visualizing multi-dimensional data
- Histograms for target data
- Gradient boosting decision trees
- Classification and regression models, feature importance
- Application of ML models in materials screening
- Application of ML in mapping energy landscape
- Web-app
- Conclusions

AI in materials

- For successful application of AI:

  -High fidelity data, pertinent algorithm and validation strategy
- Unlike other AI input data (cat/dog images from facebook etc.), materials data are really small and takes long time to generate
- Available of easily applied AI algorithms: scikit-learn, tensor-flow, lightgbm, Keras etc.
- Publicly available databases such as JARVIS-DFT, Materials-project, AFLOW, OQMD, Harvard clean energy project, AiiDA, PDB database, COD database etc.
- Immense potential for AI in materials: medicine-design, medical-diagnosis, alloy-design,…

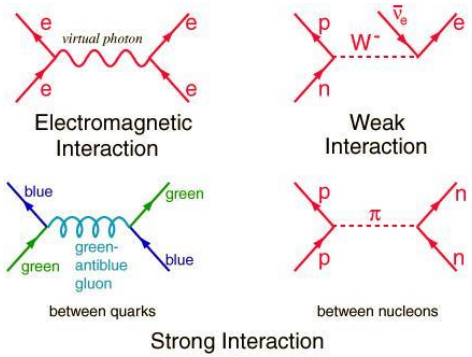


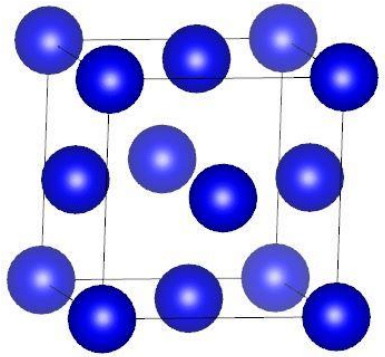


https://www.youtube.com/watch?v=8r_smb-9GTg
https://www.nesgt.com/blog/2016/10/how-artificial-intelligence-is-being-used-in-the-pharmaceutical-industry
https://rowanalytics.com/blog-post/ai-in-medicine-a-historical-perspective/

Motivation

- Bridge gap between AI and materials community
- 10^{100} materials predicted, impossible to characterize with current theoretical and experimental techniques
- Fully automated computational and experimental discovery as well as industrial implementation
- Learning data from classical-physics and quantum-physics based outputs through AI/machine learning techniques
- JARVIS : **J**oint **A**utomated **R**epository for **V**arious **I**ntegrated **S**imulations

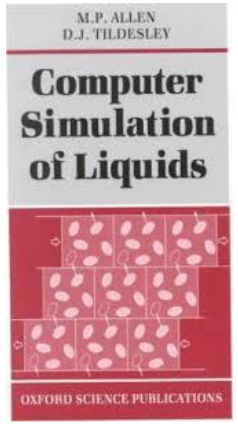




https://en.wikipedia.org/wiki/Iron_Man
https://nige.wordpress.com/2008/01/30/book/interactions/
https://arxiv.org/abs/1805.07325

JARVIS-FF

Force-field (classical)



$$F = ma = -\nabla V(r)$$

- **Solve Newton's equation for atomic positions**
- **Approximations for V (force-fields)**:

EAM, EIM, MEAM, AIREBO, REAXFF, COMB, COMB3, TERSOFF, SW *etc.*

- **Contains:**

Automated LAMMPS based force-field calculations on DFT geometries. Some of the properties included in JARVIS-FF are energetics, elastic constants, surface energies, defect formations energies and phonon frequencies of materials

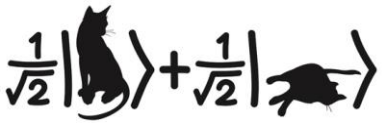
- **Time:** Takes years to fit FFs, relatively quick calculations
- **Website**: https://www.ctcms.nist.gov/~knc6/periodic.html
- **Publications**:

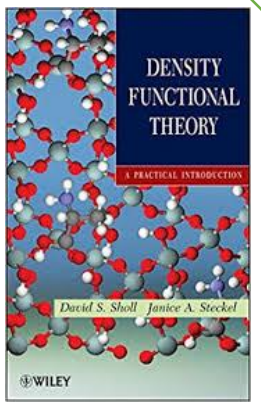
➢ Nature:Scientific Data 4, 160125 (2017)
➢ arXiv:1804.01024 (2018)

JARVIS-DFT

Density-functional theory (quantum)

Schrödinger's cat





$$H\psi = E\psi$$

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + V_{Eff}(r)\right]\psi_i(r) = E_i(r)\psi_i(r)$$

$$V_{Eff} = T + V_{Ne} + V_{ee} + V_{XC}$$

- **Solve Scrodinger equation for electrons**
- >30,000 materials data (3D, 2D, 1D, 0D)
- **Contains:**

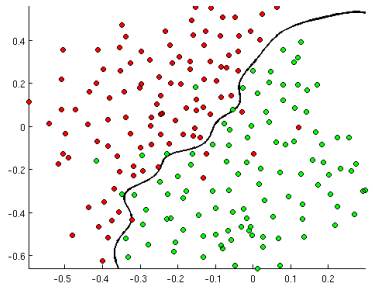
Formation energy, exfoliation energy, diffraction pattern, radial distribution function, band-structure (SOC/Non-SOC), density of states, carrier effective mass, temperature and carrier concentration dependent thermoelectric properties, elastic constants and gamma-point phonons

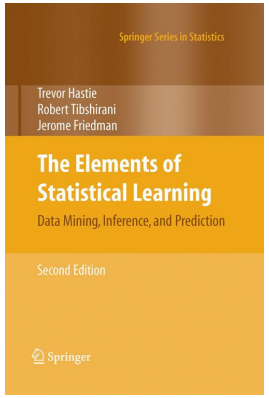
- **Time:** 5000 cores for last 4 years
- **Website**: https://www.ctcms.nist.gov/~knc6/JVASP.html
- **Publications**:

➢ Nature:Scientific Reports 7, 5179 (2017)
➢ Nature:Scientific Data 5, 180082 (2018)
➢ Phys. Rev. B 98, 014107 (2018)

JARVIS-ML

Machine learning (data-driven)





- **Drawing the line, dimensionality reduction, curve-fitting?**
- Neural nets, decision trees, fuzzy-logic etc.
- Uses gradient boosting decision tree
- **Contains:**

Machine learning prediction tools, trained on JARVIS-DFT data.

Some of the ML-predictions focus on energetics, heat of formation, GGA/METAGGA bandgaps, bulk and shear modulus, exfoliation energy, refractive index, magnetic moment, carrier effective masses

- **Time:** Much easier and faster to train
- **Website**: https://www.ctcms.nist.gov/jarvisml/
- **Publication**:

➢ Accepted Phys. Rev. Mat. (2018)

JARVIS

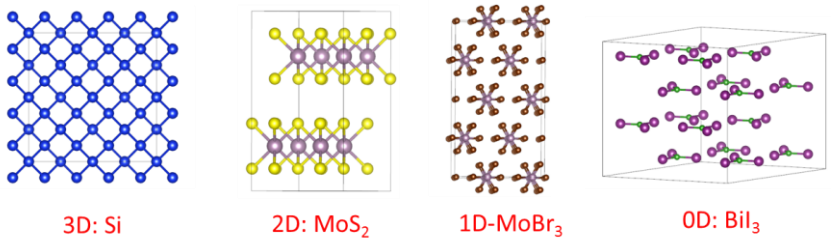
JARVIS-DFT data

- Large and reliable dataset of
  - ➢ optoelectronic properties (>18000)
  - ➢ Elastic properties (>11000)
  - ➢ 2D/1D/0D exfoliation energies (>800)
  - ➢ Topological material properties, Z_2 index (>2000)

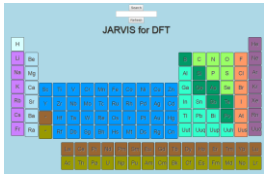
and others…

Dimensionality of materials

- **>700 2D mono/multi-layers & >30000 3D bulk materials**
- **Increasing ~3000 materials/month**

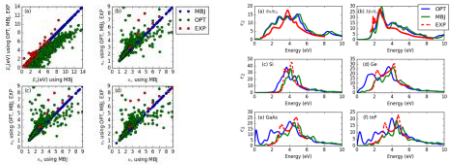


Webpage: https://jarvis.nist.gov



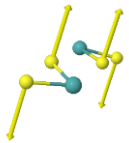
Optoelectronic properties

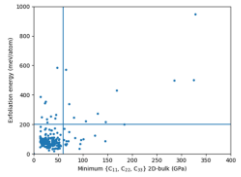
- **OptB88vdW, TBmBJ and HSE06 bandgaps**
- **Frequency dependent dielectric functions**

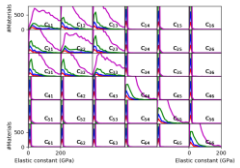


Elastic properties

- **6x6 elastic tensors, Poisson's ratio and phonons**

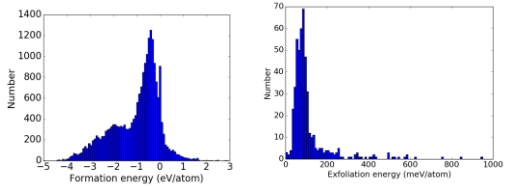






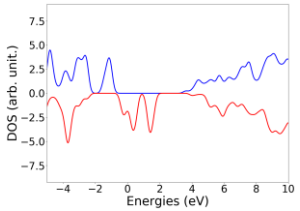
Energetics

- **Enthalpy of formation and enthalpy of exfoliation**
- **Exfoliable materials: <200 meV/atom**



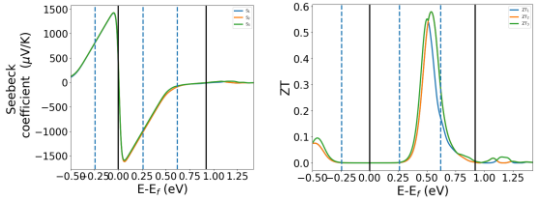
Magnetic properties

- **Magnetic moments**

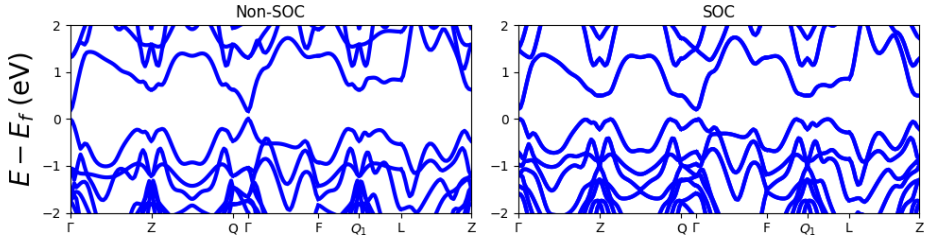


Transport properties

- **Carrier effective mass, Seebeck coefficient and zT**



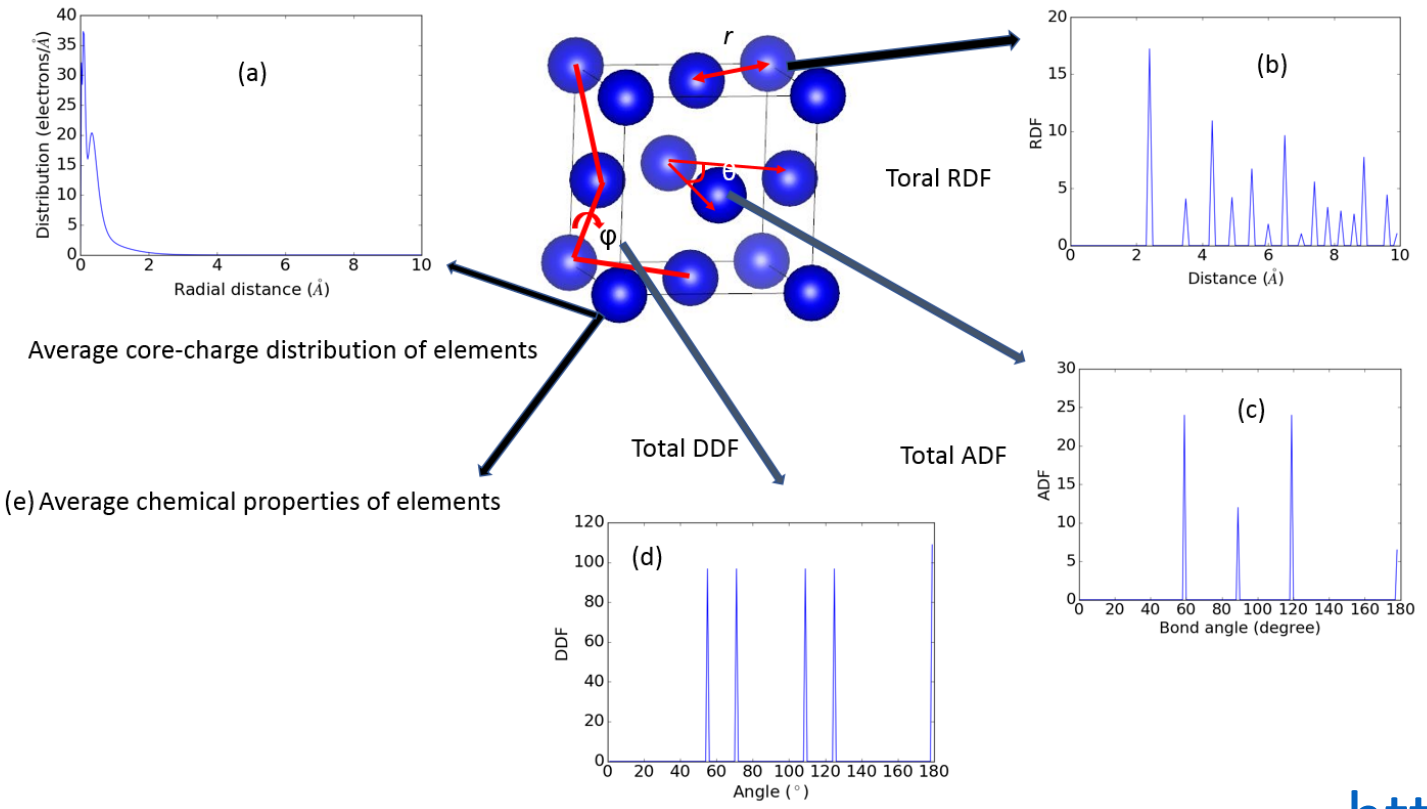
Topological properties



Steps in JARVIS-ML model

1) Descriptor selection (perhaps the most important step!),visualization
2) Preprocessing (variance threshold, PCA etc.)
3) Train-test split of data (90%-10% for instance)
4) Hyperparameter optimization using grid-search on train data
5) Select the best model (based on R^2/**MAE**/RMSE accuracy)
6) Prediction on test data (classification/regression)
7) Plot learning curve for complexity analysis
8) Plot feature importance (if available)

Finding the right features/descriptors: representing atomic structure to computers

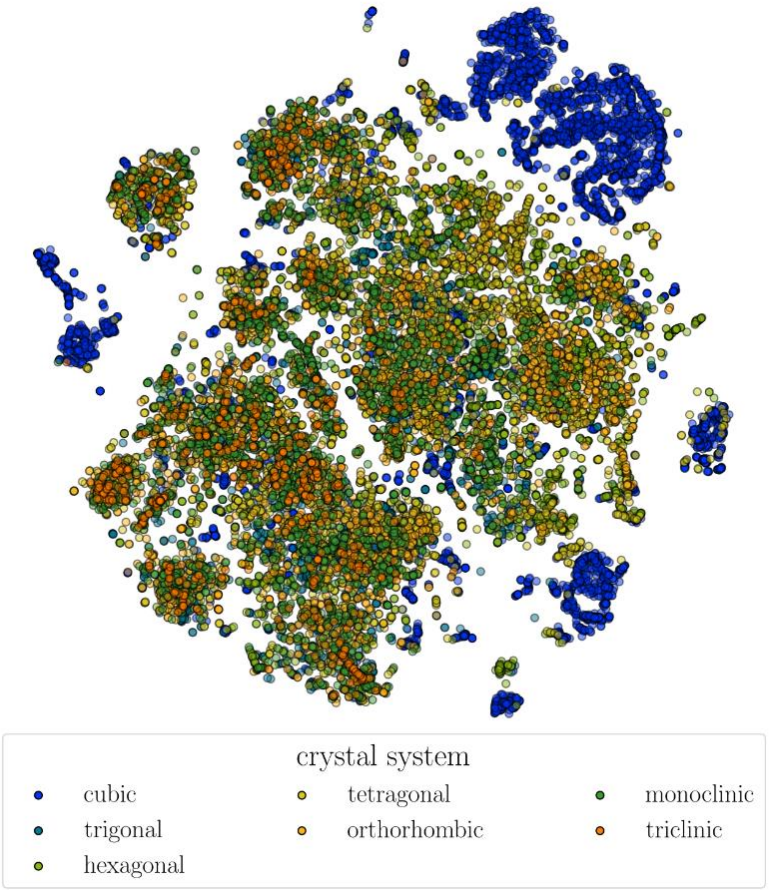


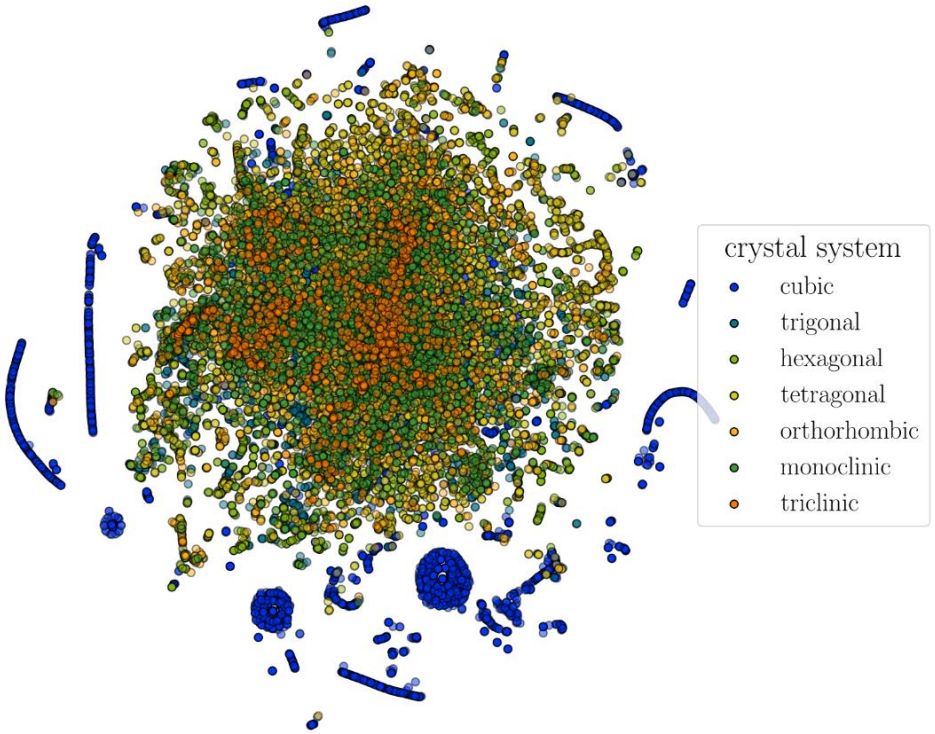
- Arithmetic operations (mean, sum, std. deviation…) of **electronegativity, atomic radii, heat of fusion**,….
of atoms at each site
(example: Electronegativity of Mo+Mo+S+S+S+S)/6 = 0.15
- Atomic bond distance based descriptors
- Angle based descriptors

1557 descriptors/features for one material

https://github.com/usnistgov/jarvis

Visualizing multi-dimensional data with t-SNE



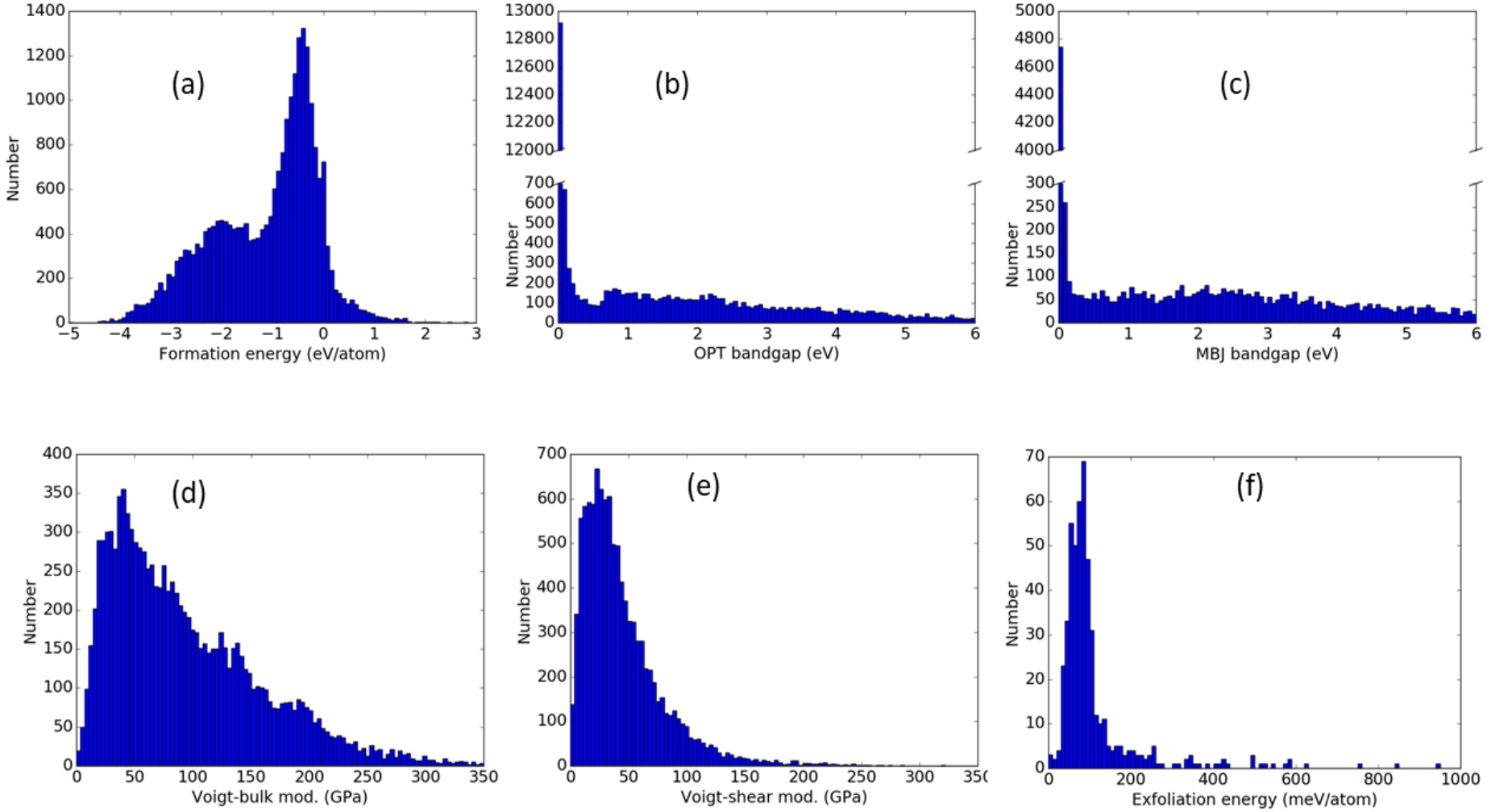


- Converts similarities between data points to joint probabilities
- Visulaization with t-SNE for ~25000 materials

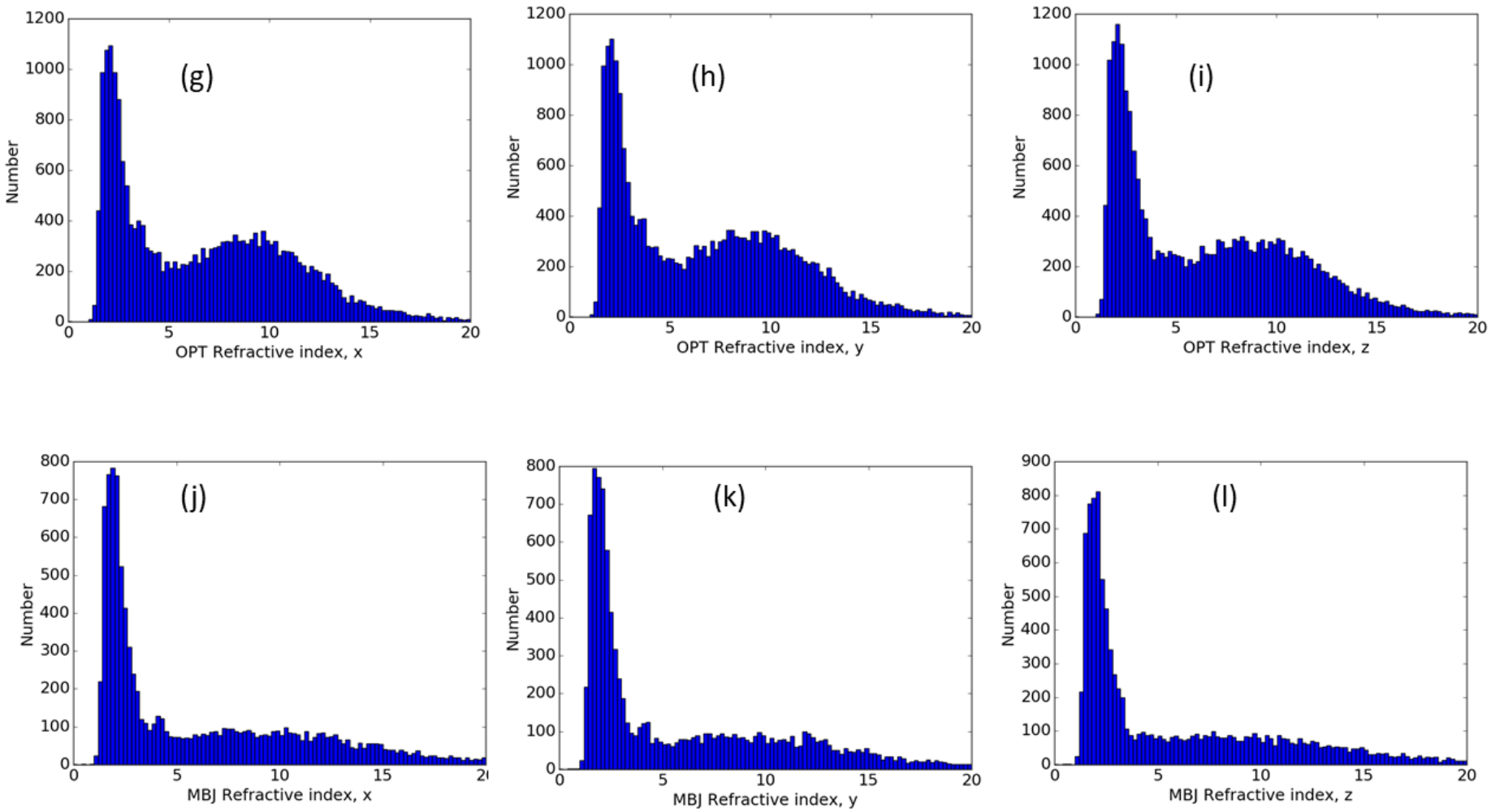
http://holoviews.org/
http://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html

Data-spread: histograms

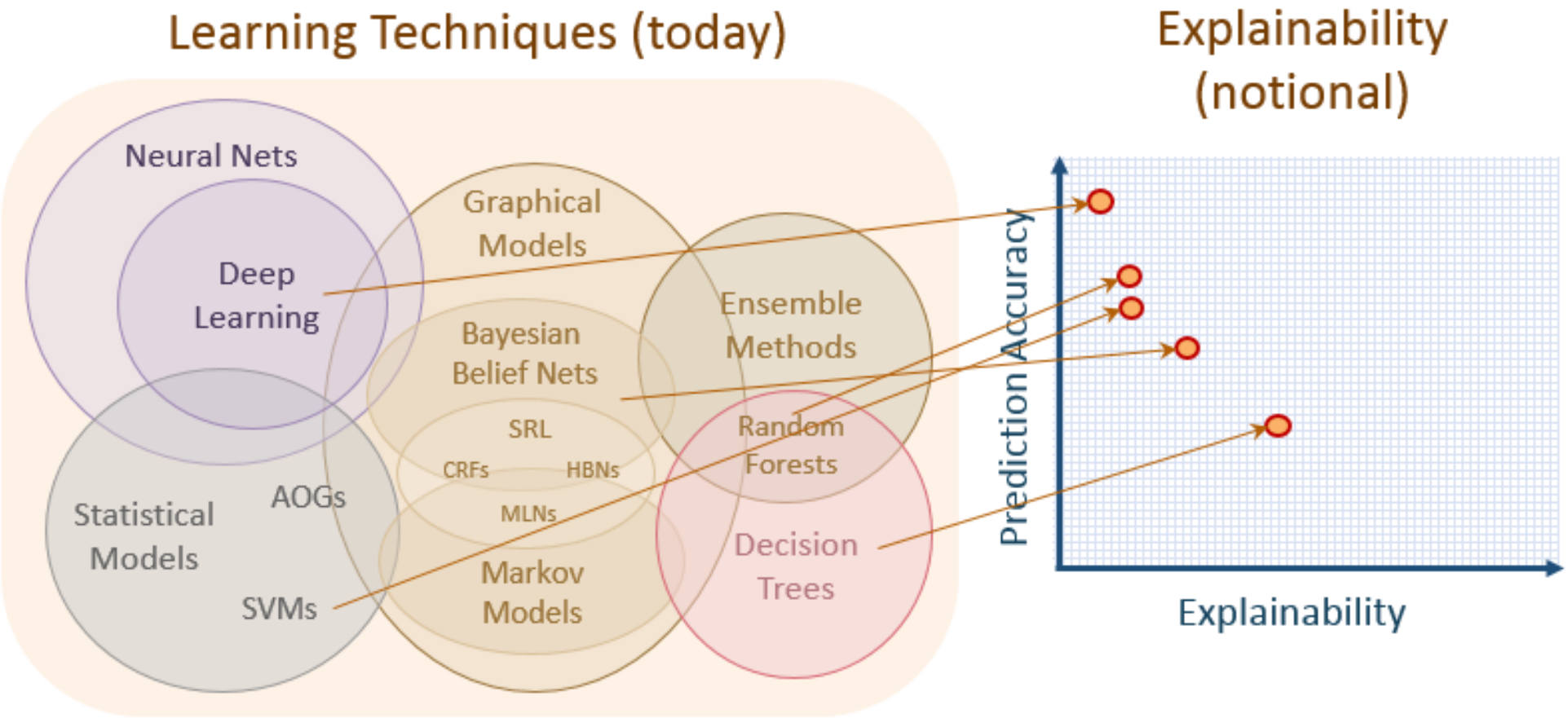
- Identifying the range of target data
- Energetics, electronic, optical, and mechanical properties
- AI generally good for interpolation



Data-spread: histograms



Need of explainable AI for materials: black box may not work for materials community



https://www.darpa.mil/program/explainable-artificial-intelligence

Gradient boosting decision tree (GBDT)

- Ensemble of weak decision tree models
- Consecutively fits new models to provide a more accurate estimate of the response variables
- Suppose there are *N* training examples: $\{(x_i, y_i)\}^N$
- GBDT model estimates the function of future variable *x* by the linear combination of the individual decision trees using:

$$f_m(x) = \sum_{m=1}^{M} T(x; \theta_m)$$

where $T(x; \theta_m)$ is the *i*-th decision tree, θ_m is its parameters and *M* is the number of decision trees

- Final estimation in a forward stage-wise fashion

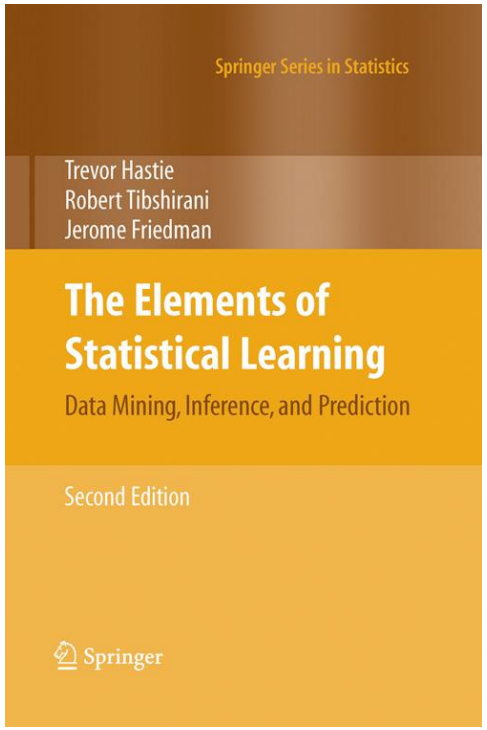
$$f_m(x) = f_{m-1}(x) + T(x; \theta_m)$$

where $f_{m-1}(x)$ is the model in $(m-1)$ step. The parameter θ_m is learned by the principle of

- Empirical risk minimization using:

$$\widehat{\theta_m} = arg\,\underset{\theta_m}{min} \sum_{i=1}^{N} L\big(y_i, f_{m-1}(x) + T(x; \theta_m)\big)$$

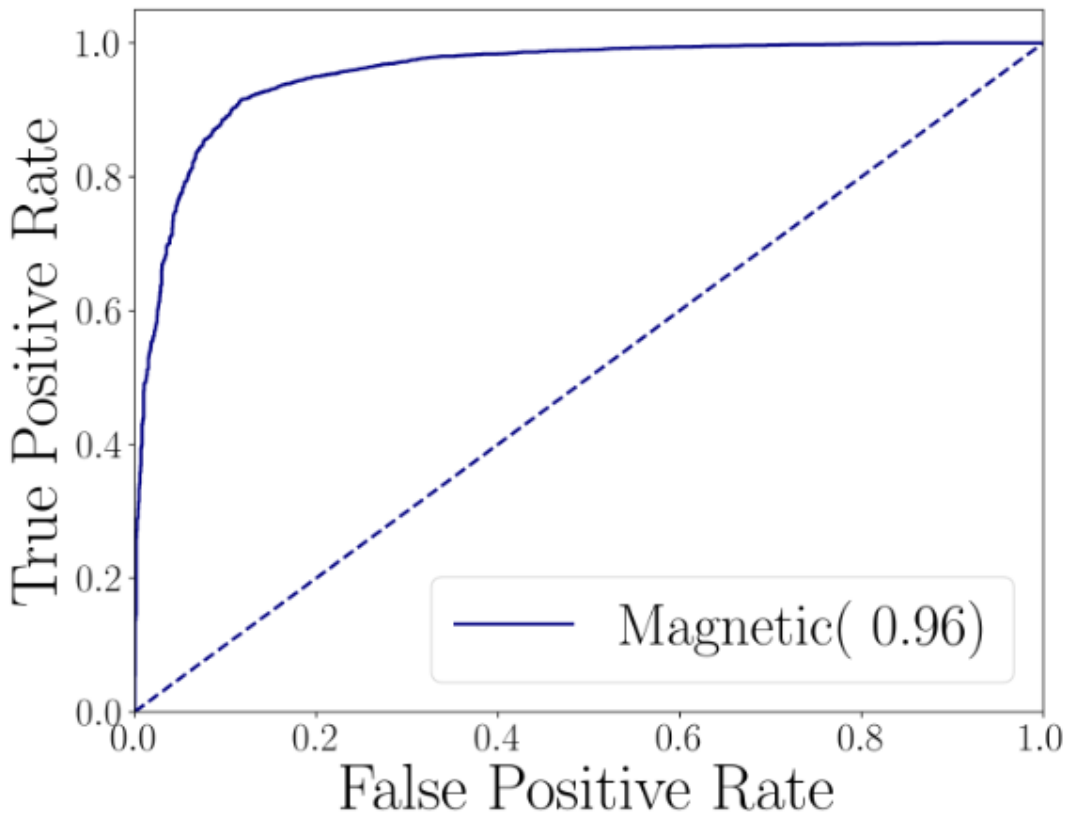
where *L* is the loss-function.

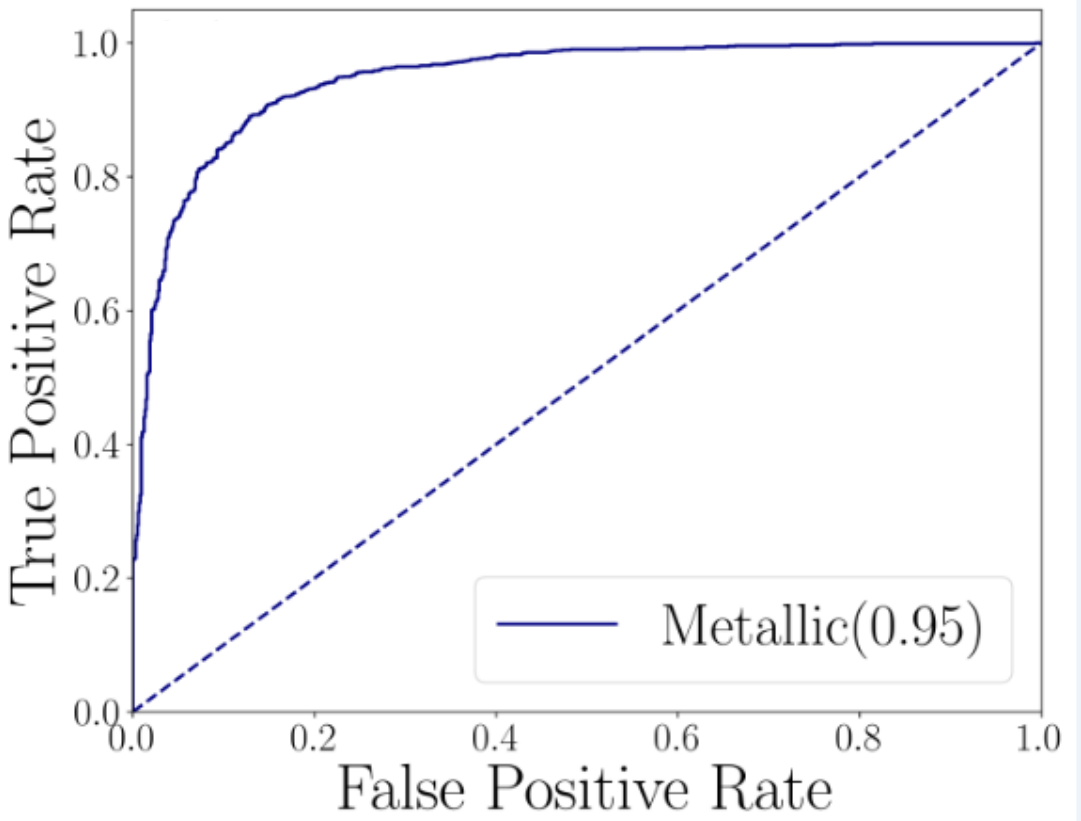


https://web.stanford.edu/~hastie/Papers/ESLII.pdf

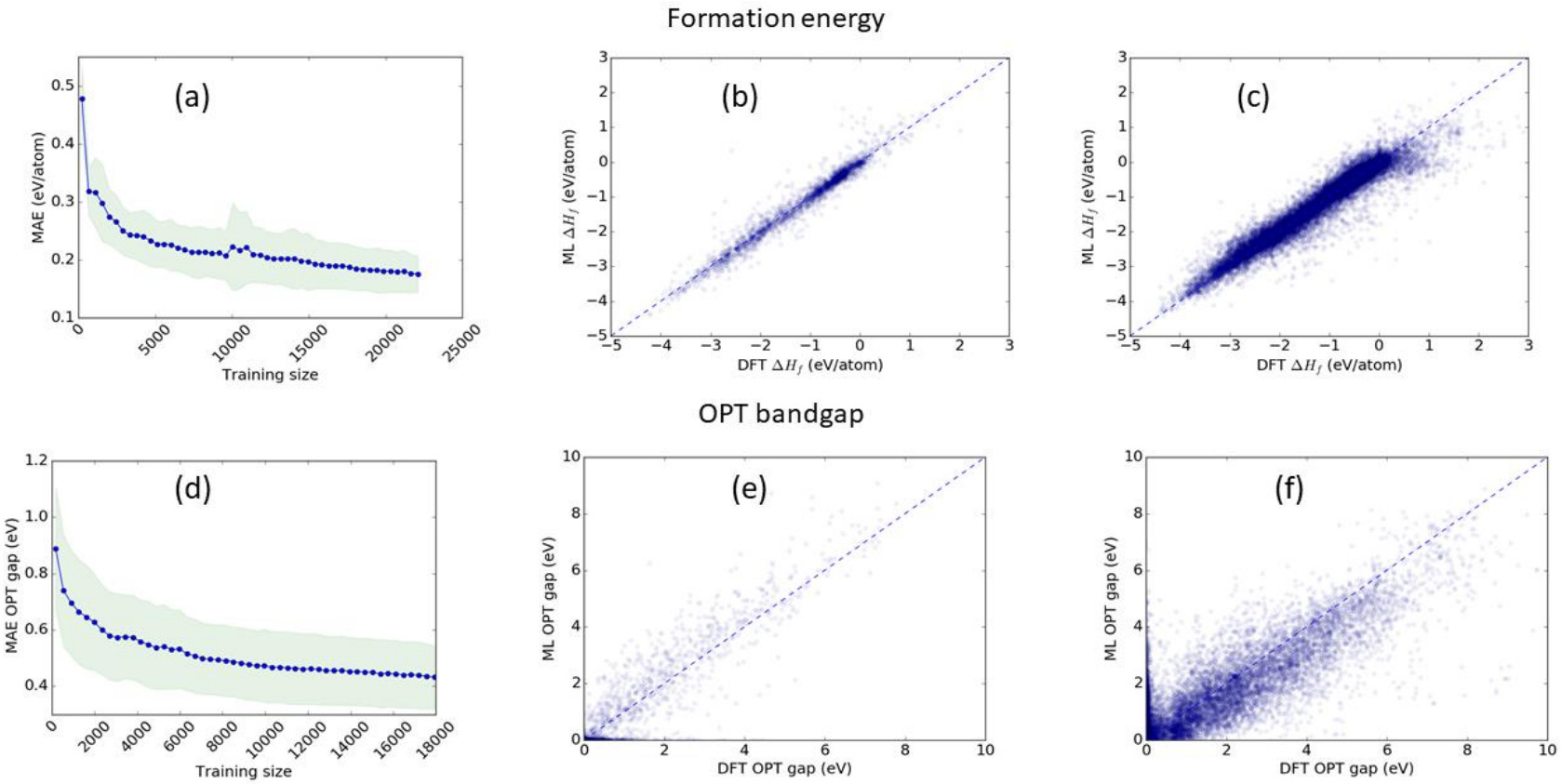
Classification problem: metal/non-metal, magnetic/non-magnetic materials

ROC-curve: Excellent classification models





Regression models: formation energy and bandgap model



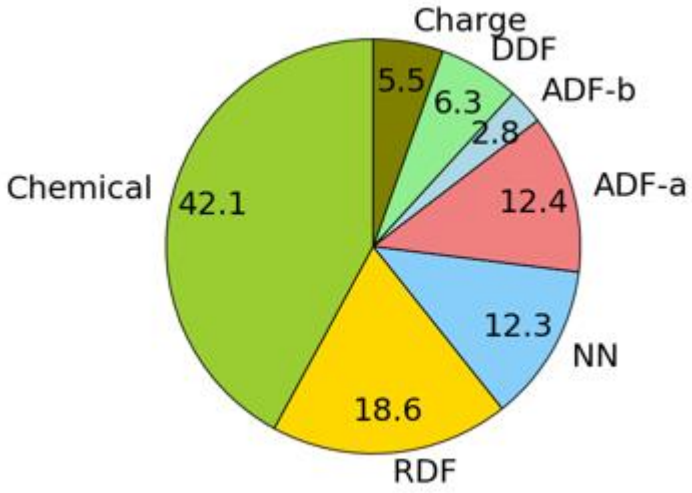
Learning curve shows scope of further improvement

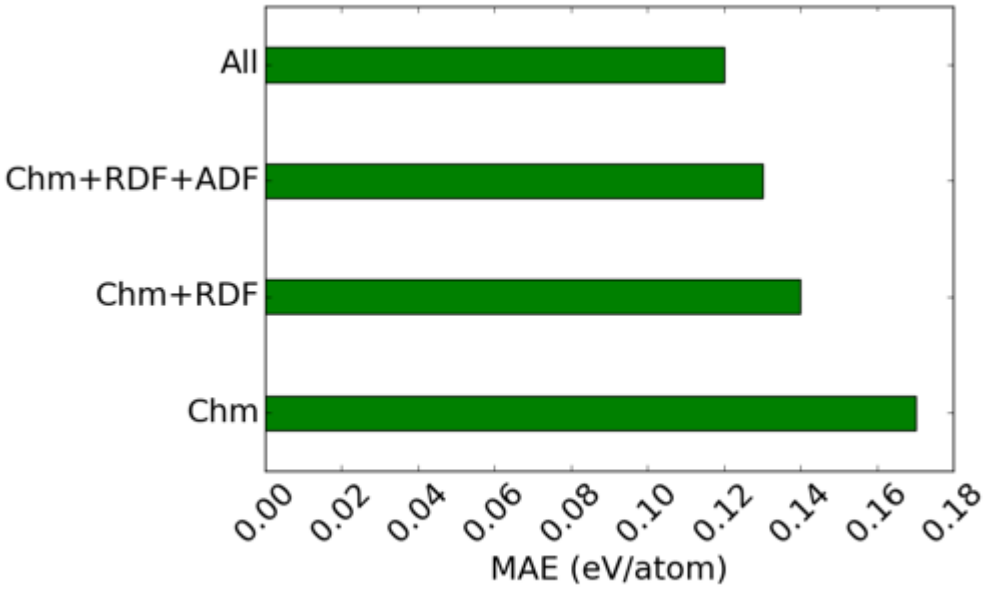
Other model performance

Property	#Data-points	$MAE_{CFID-DFT}$	$MAE_{DFT-Exp}$
Formation energy (eV/atom)	24549	0.12	0.136
Exfoliation energy (meV/atom)	616	37.3	-
OPT-bandgap (eV)	22404	0.32	1.33
MBJ-bandgap (eV)	10499	0.44	0.51
Bulk modulus (GPa)	10954	10.5	10.0
Shear modulus (GPa)	10954	9.5	10.0
OPT-n_x (no unit)	12299	0.54	1.78
OPT-n_y (no unit)	12299	0.55	-
OPT-n_z (no unit)	12299	0.55	-
MBJ-n_x (no unit)	6628	0.45	1.6
MBJ-n_y (no unit)	6628	0.50	-
MBJ-n_z (no unit)	6628	0.46	-

Performance on 10 % held data

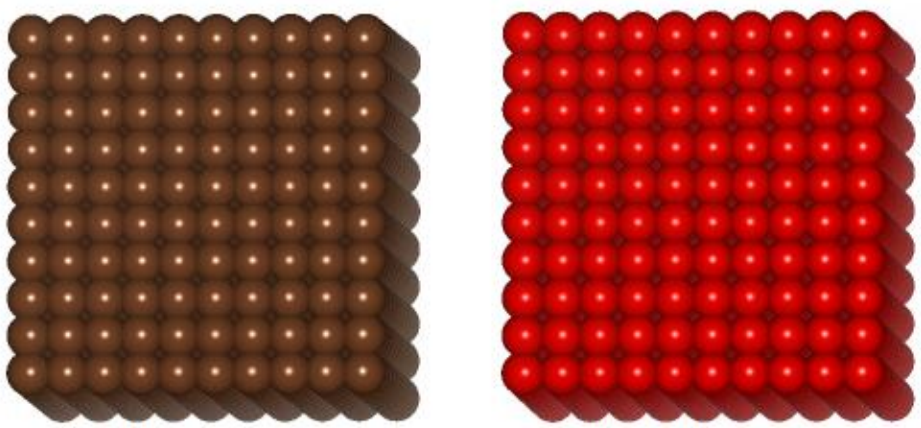
Explainability: feature importance





- Chemical features most important followed by RDF and NN
- Incrementally adding structural features decreases MAE

Introduction to Genetic Algorithm



(a) Artificial cubic parent structure. (b) Artificial cubic parent structure.

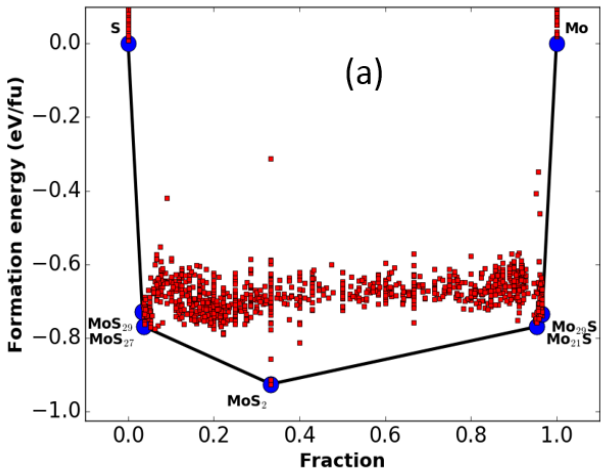
(c) Child created by the slicing crossover using a horizontal cut. (d) Child created by the slicing crossover using a periodic cut.

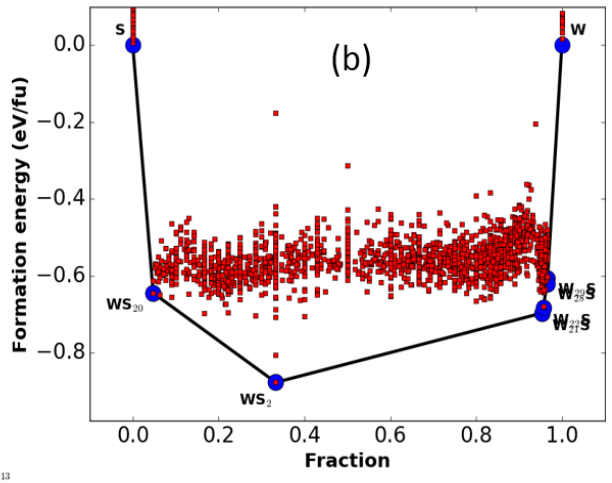
Picture from GASP manual

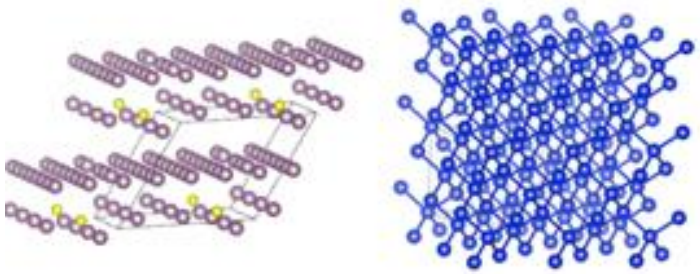
- Based on **'Survival of the fittest'** theory: fitness of crystal structure based on energy of structure
- Parents to offspring crystal structure
- Generally energy is obtained from DFT, MD…let's try ML ?

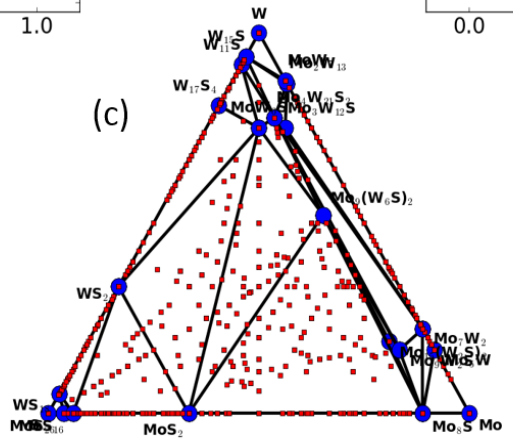
D. M. Deaven, Molecular geometry optimization with a genetic algorithm, Physical Review Letters, 75 (1995)
G. Ceder, Data-mining-driven quantum mechanics for the prediction of structure, MRS Bulletin, 31 (2006)
https://github.com/henniggroup/GASP-python/

Search for new materials: Genetic algorithm with ML





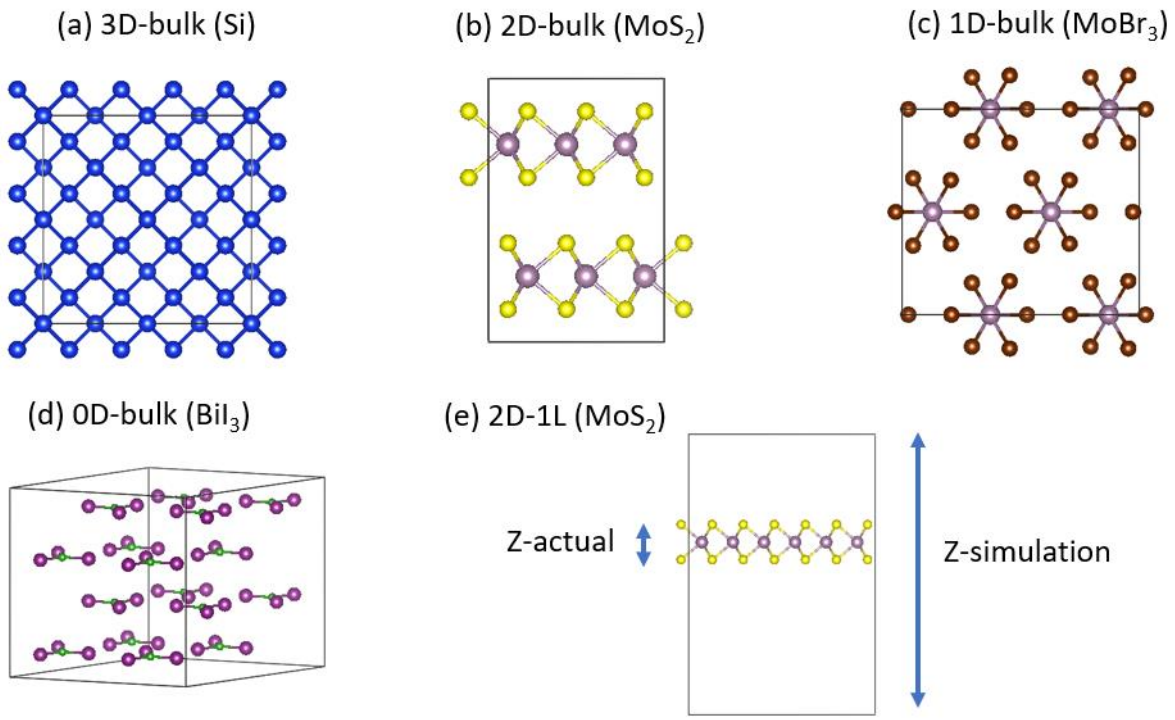




- New way of validating ML model for materials

- MoS_2, WS_2 indeed stable as in DFT and experiments
- Need further verification foor low-lying energy structures with DFT

2D materials screening: flexible electronics applications

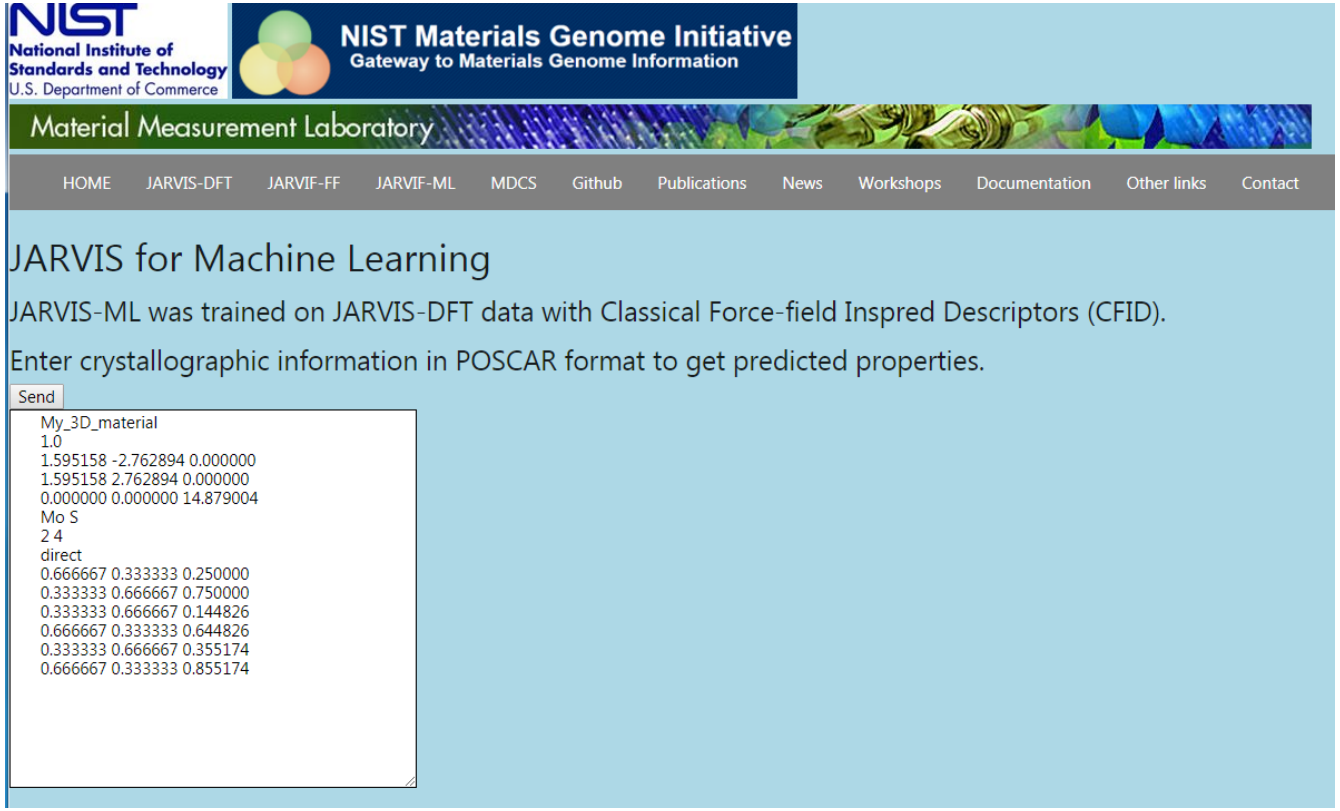


- ~5000 2D materials predicted
- Requires expensive DFT calculations for predicting properties such as bandgap, exfoliation energy etc.
- Use of ML drops down the time to a few seconds
- Using this technique we identified new 2D materials such as CuI, InS etc.
- Validated using DFT

Web-app: DEMO

https://www.ctcms.nist.gov/jarvisml/

- Pickle the learnt parameters
- Integrate with FlaskPython



Getting hands on: github sample trainng

https://github.com/usnistgov/jarvis/blob/master/jarvis/db/static/jarvis_ml-train.ipynb

https://github.com/usnistgov/jarvis/blob/master/jarvis/sklearn/examples/desc_example.py

usnistgov / jarvis

Unwatch 7 | Star 1 | Fork 6

Code | Issues 0 | Pull requests 0 | Projects 0 | Wiki | Insights | Settings

Branch: master | jarvis / jarvis / db / static / jarvis_ml-train.ipynb | Find file | Copy path

JARVIS-Unifies jarvis-ml nb 75836d0 24 seconds ago

1 contributor

1054 lines (1053 sloc) 110 KB | Raw | Blame | History

Training a JARVIS-ML model

In [64]:

```
#from jarvis.sklearn.get_desc import get_comp_descp
from monty.serialization import loadfn, MontyDecoder,dumpfn
import numpy as np
from sklearn.metrics import mean_absolute_error,r2_score,mean_squared_error
import lightgbm as lgb
import matplotlib.pyplot as plt
plt.switch_backend('agg')
%matplotlib inline
import pandas as pd
from sklearn.datasets import load_boston
from sklearn.model_selection import train_test_split,learning_curve,cross_val_score,cross_val_predict,Grid
SearchCV,RandomizedSearchCV
import scipy as sp
import time,os,json,pprint
from sklearn.feature_selection import SelectKBest,f_classif,SelectFromModel,VarianceThreshold
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import StandardScaler


# Download descriptors and material data from the following link:
#  https://figshare.com/articles/JARVIS-ML-CFID-descriptors_and_material_properties/6870101

# Websites: https://www.ctcms.nist.gov/jarvisml, https://jarvis.nist.gov
# https://arxiv.org/abs/1805.07325


# NIST-disclaimer: https://www.nist.gov/disclaimer
```

On-going work

- Combining all materials data >5 million:
  solid-state crystals, molecules, proteins in one database and their visualization
- Multi-output Regression
- Active learning: train on DFT data, add to experiments to reduce domain search
- Transfer learning: Use previously trained model to learn on new data

Conclusions

- Golden time to integrate physics and data-science
- ML/AI as an aid to conventional theoretical methods such as DFT
- JARVIS bridging the gap between data-science and physics based models
- Providing unique material-descriptors
- All the code and data publicly available
- Formation energy convex hull plot as an example, other multicomponent systems also possible
- Web-app for on-the fly prediction of properties
- Immense potentials for electronics and biomedical industry
- Important links:

✓ https://jarvis.nist.gov/

✓ https://www.ctcms.nist.gov/jarvisml/

- E-mail: kamal.choudhary@nist.gov