Machine Learning for Materials Science

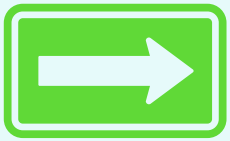
Prasenjit Sen
Prof. (HRI)

Major areas of use

Predicting materials properties
Using trained ML models

Searching materials space
using generative AI

ML force fields for fast
Simulations at quantum accuracy



Accelerating materials characterization
Fast interpretation of XRD, XPS spectra

What are property prediction models?

- Regression models
  - Predicting properties that take continuous values
    - Formation energy (or heat of formation h_{form}), band gap, saturation magnetization (M_s), elastic constants, electronic thermal conductivity, lattice thermal conductivity.

- Classification models
  - Models for finding class/category of materials
    - Stable or unstable? Magnetic or non-magnetic? Ferromagnetic or anti-ferromagnetic? All examples of binary classification.

- Find hidden pattern in data: unsupervised model, clustering ‘similar’ materials; ideas of graph theory may be useful.

What can we do with property prediction models?

Screening materials

- Experiments: Very expensive, time-consuming, uncertain
  - High throughput experiments, expensive still, uncertain
- Predict material properties from theory
  - DFT+ methods, High throughput: better, still expensive
- Can we bypass the calculations?

Leverage available data for Machine Learning

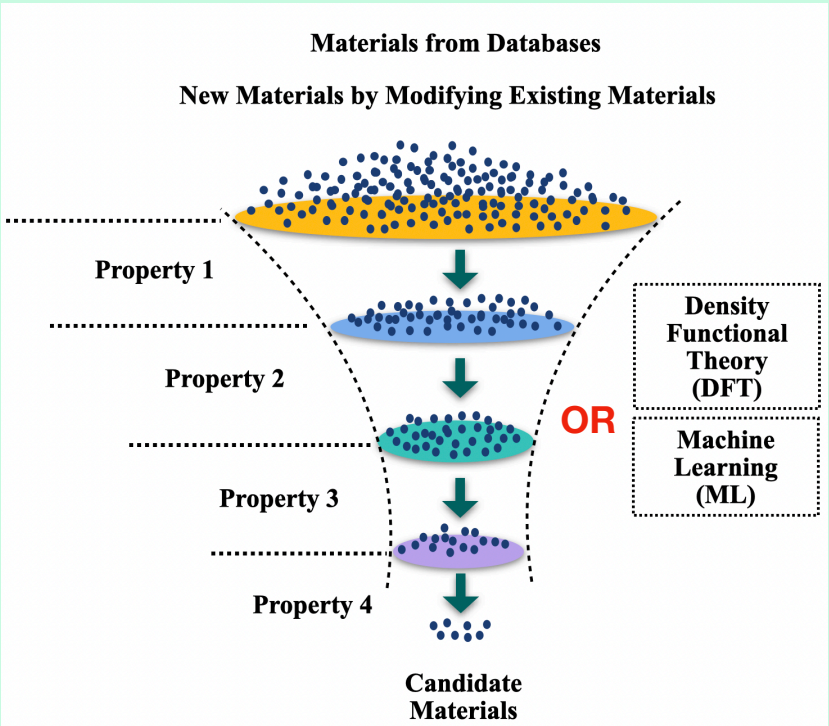
Slow pace of materials discovery and deployment

Trial-and-error experimentation

Giant magnetoresistance materials (1988)
Data-storage storage (1997)

Li-ion battery tested in 1970s
Widely used only in 1990s
Still some way before widely used for mobility

Materials Genome Initiative (MGI, 2011)
Envisaged accelerated materials discovery and deployment



Train ML models to predict properties

Leveraging available data

Model training and performance

Material property

$$Y = f(\vec{X}) + \epsilon = f(X_1, X_2, \ldots, X_n) + \epsilon$$

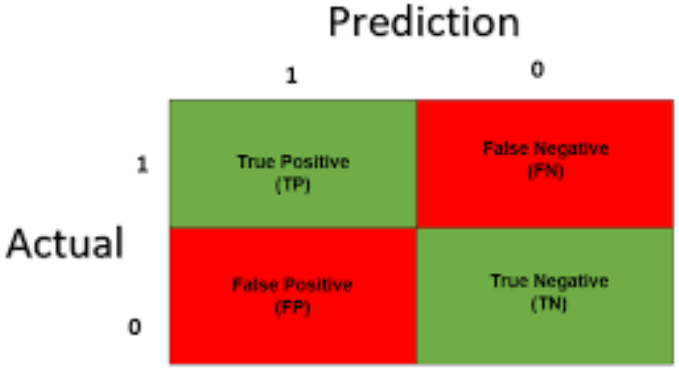
We get the best estimate

$$\hat{Y} = \hat{f}(X_1, X_2, \ldots, X_n) + \epsilon$$

- **Regression model performance metrics**
  - Quality of fit, $R^2 = 1 - \frac{\sum_i [Y^i - \hat{Y}^i]^2}{\sum_i [Y^i - \bar{Y}^i]^2}$
  - Correlation $r = \langle Y^i, \hat{Y}^i \rangle$
  - Mean absolute error, MAE $= \frac{1}{N} \sum_i | Y^i - \hat{Y}^i |$
  - Mean square error

    MSE $= \frac{1}{N} \sum_i (Y^i - \hat{Y}^i)^2$

Confusion matrix



- **Classification model performance metrics**
  - Accuracy $= \frac{TP + TN}{N}$
  - Precision $= \frac{TP}{TP + FP}$
  - Recall $= \frac{TP}{TP + FN}$
  - f_1 score $= \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$

Example

Confusion matrix for a model predicting loan default

	True default status			
Predicted default status		Yes	No	**Total**
	Yes	81	23	104
	No	**252**	9644	9896
	Total	333	9667	10000

$$\text{Accuracy} = \frac{81 + 9644}{10000} = 0.97$$

$$\text{Precision} = \frac{81}{104} = 0.78$$

$$\text{Recall} = \frac{81}{333} = 0.24$$

$$f_1 = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} = 0.37$$

- **Classification model performance metrics**
  - $\text{Accuracy} = \frac{TP + TN}{N}$
  - $\text{Precision} = \frac{TP}{TP + FP}$
  - $\text{Recall} = \frac{TP}{TP + FN}$
  - $f_1 \text{ score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$

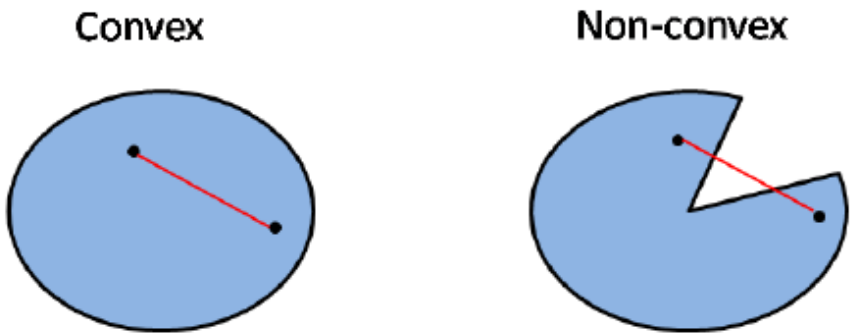
Definitions

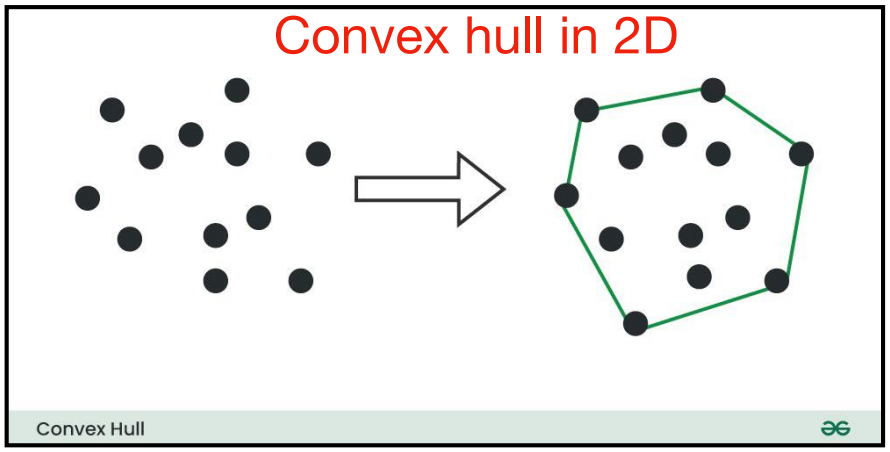
- **Formation energy** or **Heat of formation**
- $$h_{\text{form}} = E(\text{material}) - \sum_i \mu_i(\text{constituents in ref state})$$
  - $h_{\text{form}}(\text{ABX}_3) = E(\text{ABX}_3) - \mu_{\text{A}} - \mu_{\text{B}} - 3\mu_{\text{X}}$, extensive quantity
  - μ's are the chemical potentials of the constituents in their 'reference states'
  - Reference state in a particular case depends on the conditions of synthesis
  - If we take ref states to be isolated atomic states, we get cohesive energy E_{coh}
  - Can also be other states (bulk solid, molecular gas etc.)

- Energy convex hull

  Convex hull is the smallest **convex set** that encloses all the points, forming a convex polygon.

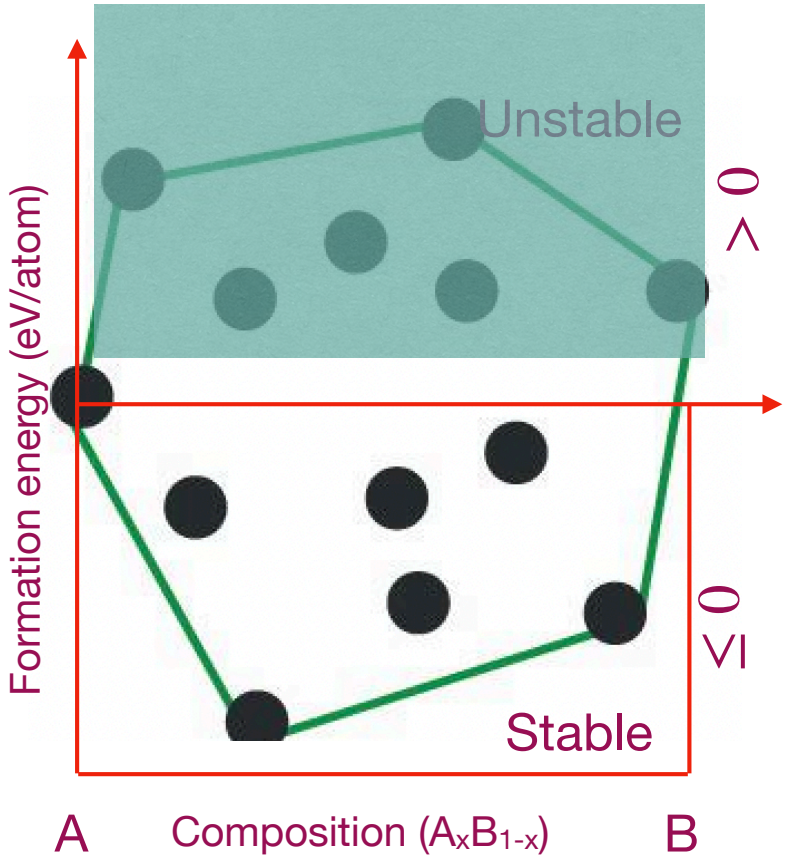
A set $\mathbf{P}$ is **convex** if for any $p, q \in \mathbf{P}$, the segment pq lies entirely in $\mathbf{P}$.



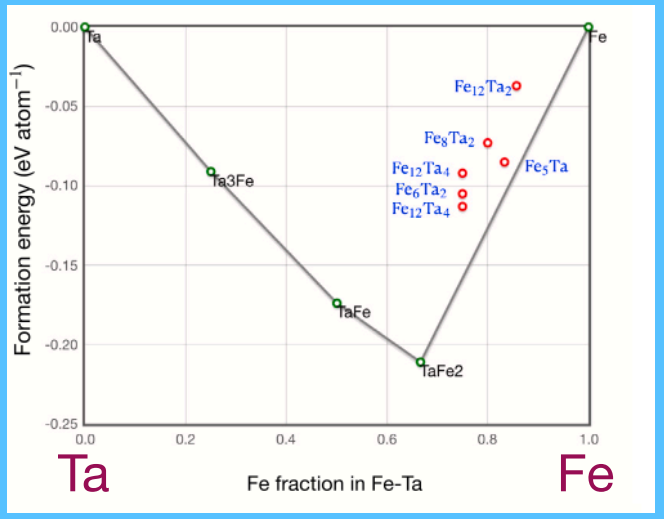


More examples:
https://in.mathworks.com/help/matlab/math/types-of-region-boundaries

Convex hull in binary phase space



Convex hull in the Fe-Ta phase diagram

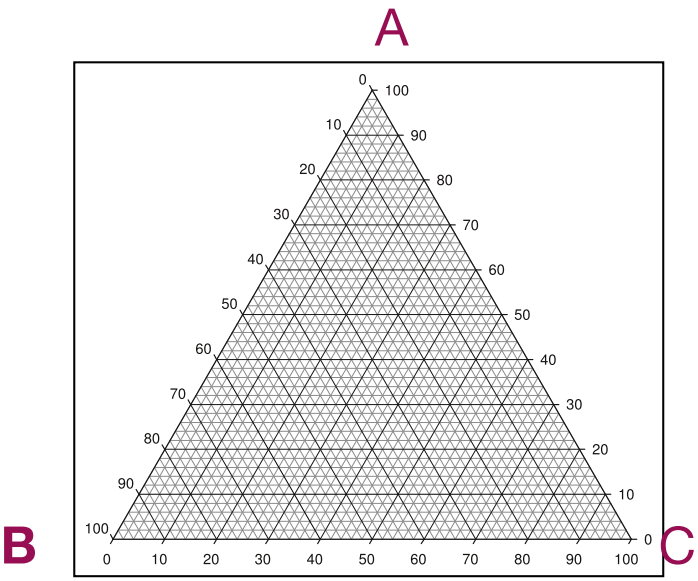


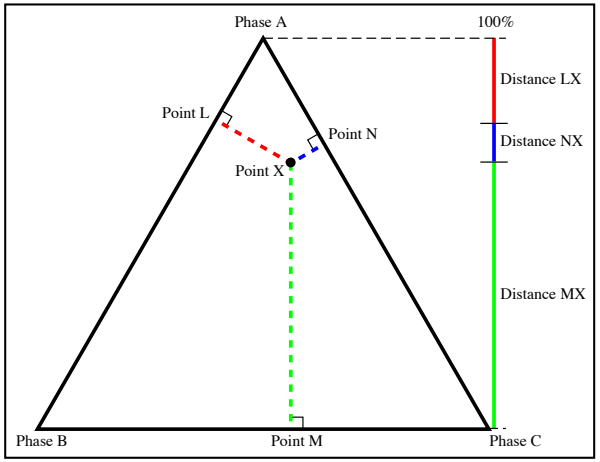
S. Mal & PS, JMMM (2024)

Convex hull in ternary phase space

Understanding the representation

- Equilateral triangle
- Three corners: three elemental phases
- Sides: binary phases
- Points in the triangle: ternary phases
- Three ways to measure composition





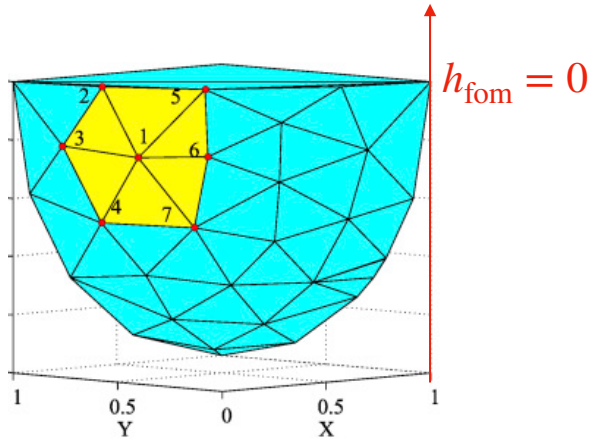
Axis perpendicular to the plane of the triangle, plotting h_{form}.

We are interested in materials with $h_{\text{form}} < 0$

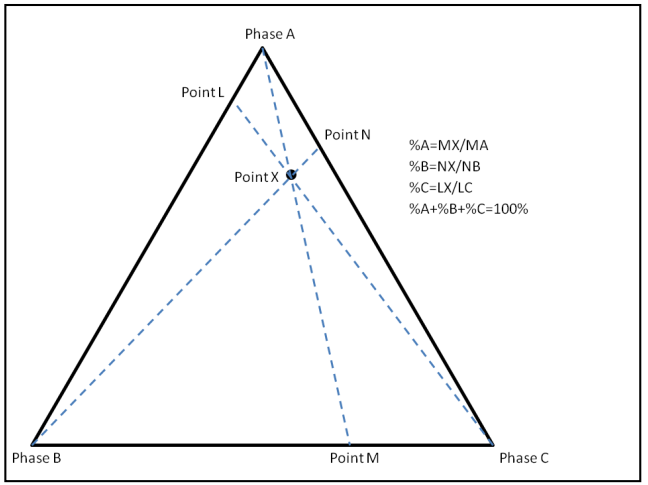
The planes (the small triangles in the figure), forming the outermost surface, constitute the convex hull

Materials on the hull most stable at that composition

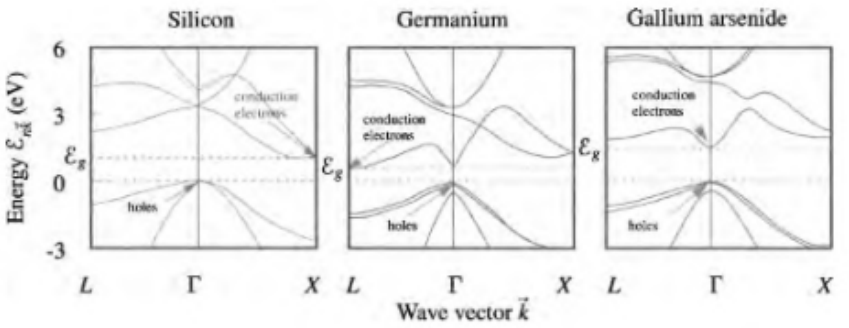
Distance from the hull along h_{form} axis a measure of instability



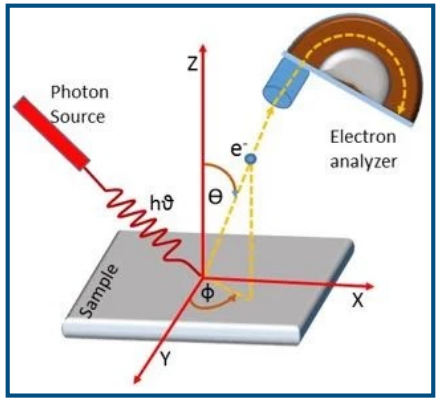
(a) A convex hull with all vertices



- **Band gap** for semiconductors/insulators in eV



- Fundamental gap, $E_g^{\min} = min\{[E(N+1) - E(N)] - [E(N) - E(N-1)]\}$
  - Gap measured via ARPES is $E_g^{\min}$



ARPES setup

  - Transport gap:

$$\sigma(T) \sim e^{-E_g/2K_BT}; \quad \ln(\sigma) \sim -E_g^{tr} + \frac{1}{2K_BT}.$$

  - Optical gap, Tauc plots

  - Gaps calculated via DFT (LDA, GGA, HSE etc.)
  - Gaps calculated via many-body methods such as GW

Do not mix different band gaps in your training data!!

- **Magnetization** in ferromagnet: Magnetic moment when all moments in a material point in the same direction
  - Units of saturation magnetization density in atomistic calculations

    M in $\mu_B/\text{Å}^3$
  - Expressed in Tesla in practical situations
  - In Tesla, $M_s = \mu_0 M = 11.649$ T
    - μ_0 is permeability of free space, $\mu_0 = 4\pi \times 10^{-7}$ N/Å^2.