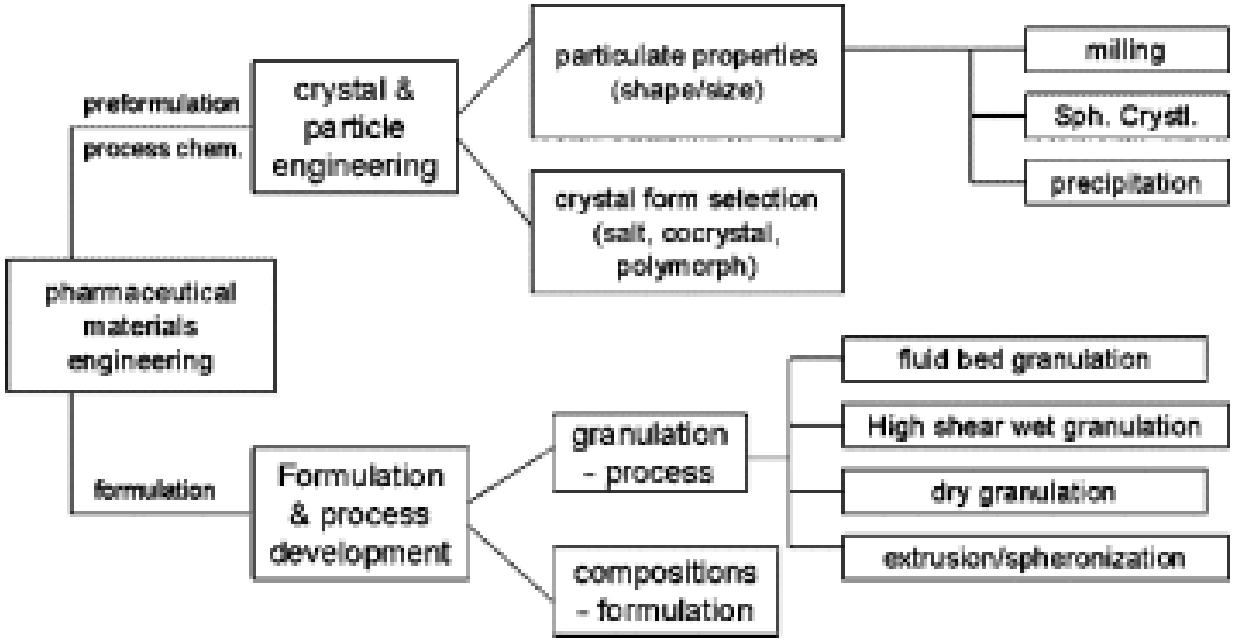
REPRESENTATIONS FOR (INORGANIC) MATERIALS

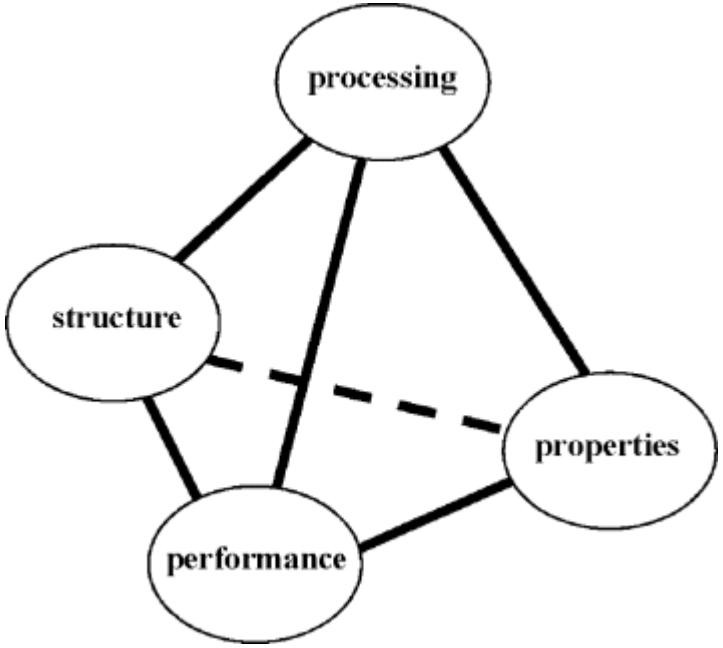
Logan Ward
Asst. Computational Scientist
Argonne National Laboratory

29 January 2021

What sets “materials” apart?

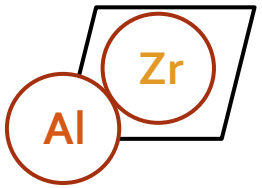
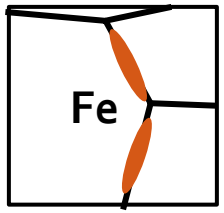
Relationships between “processing,” “structure,” and “properties”





Ref: Sun. J. Phar. Sci., (2009)

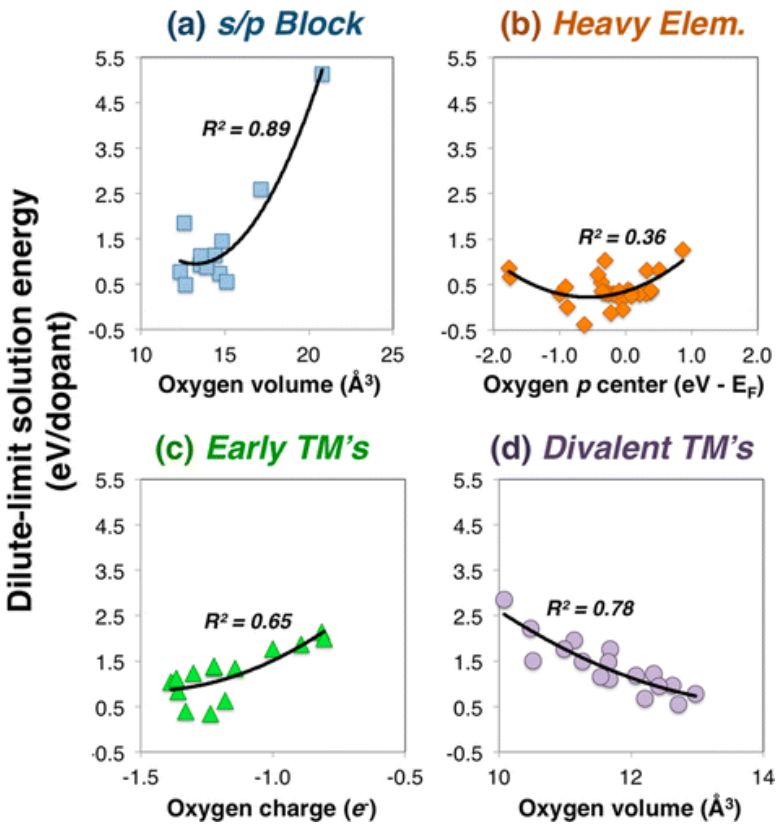
Complication: Materials are multi-scale

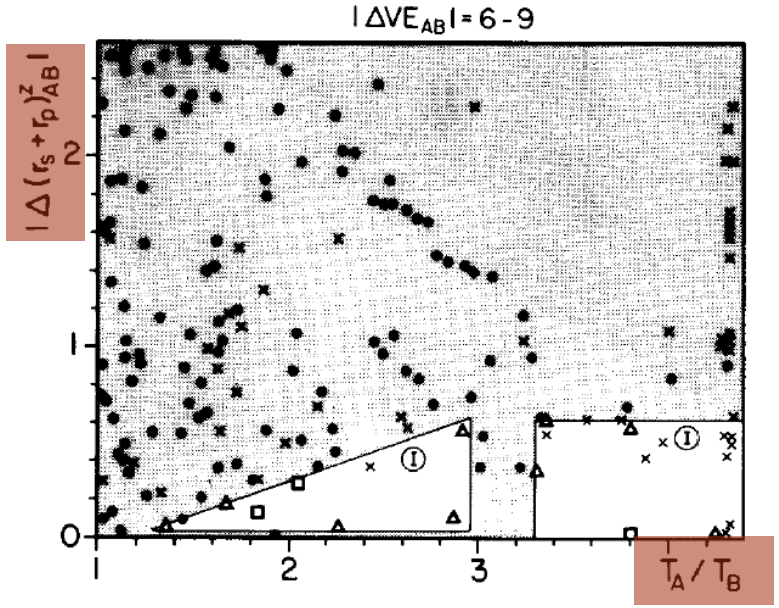
	Zr	Al_3Zr	Al_4Zr		
Available Information	Element	Phase Diagram	Composition	Crystal Structure	μ-Structure
Example Properties	Dilute ΔH_{mix}	# Eutectics	Glass-Forming Ability	ΔH_f	σ_Y
Common Representations	Element Property	Elemental Properties	Elemental Properties	Local Environment	n-Body Statistics

Element-property based features are very common

Element → **Phase Diagram** → **Composition**

Element Properties → *Property Differences* → *Property Statistics*





$$x_{\text{avg}} = \frac{1}{\sum_{i=1}^{N} \alpha_i} \sum_{i}^{N} \alpha_i x_i \tag{5}$$

$$x_{\text{red}} = \frac{1}{(N-1)\sum_{i=1}^{N} \alpha_i} \sum_{i \neq j}^{N} \left(\alpha_i + \alpha_j\right) \frac{x_i x_j}{x_i + x_j} \tag{6}$$

$$x_{\text{diff}} = \frac{1}{(N-1)\sum_{i=1}^{N} \alpha_i} \sum_{i \neq j}^{N} \left(\alpha_i + \alpha_j\right) |x_i - x_j| \tag{7}$$

where when considering a compound, $A_a B_b C_c$, we define $\boldsymbol{\alpha}$ as the vector of coefficients $[a, b, c]$ and N as the length of $\boldsymbol{\alpha}$. For example, for $CaTiO_3$, $\boldsymbol{\alpha} = [1,1,3]$ and $N = 3$.

Ref: Meredig and Wolverton. Chem. Mat. (2014)

Ref: Villars. JLCM. (1985)

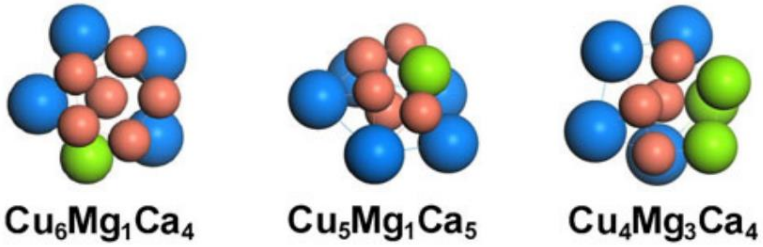
Ref: Bartel et al. Nat. Comm. (2018)

There is also much room for creativity

Intuition-based features

(Similar spirit to chemical descriptors)

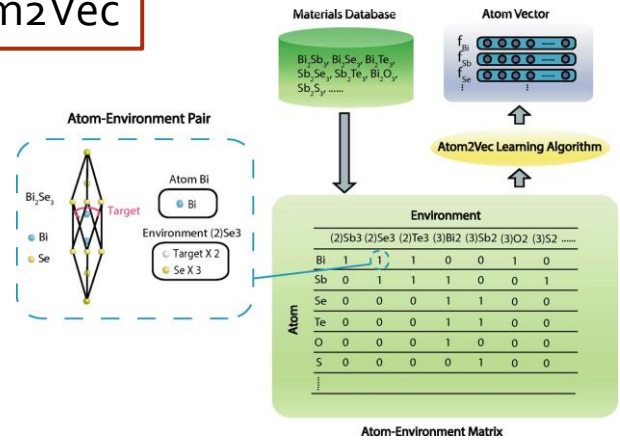
Cluster Packing Efficiencies: Laws *et al.* MMTA (2010)



Many available in matminer

Learning from Data

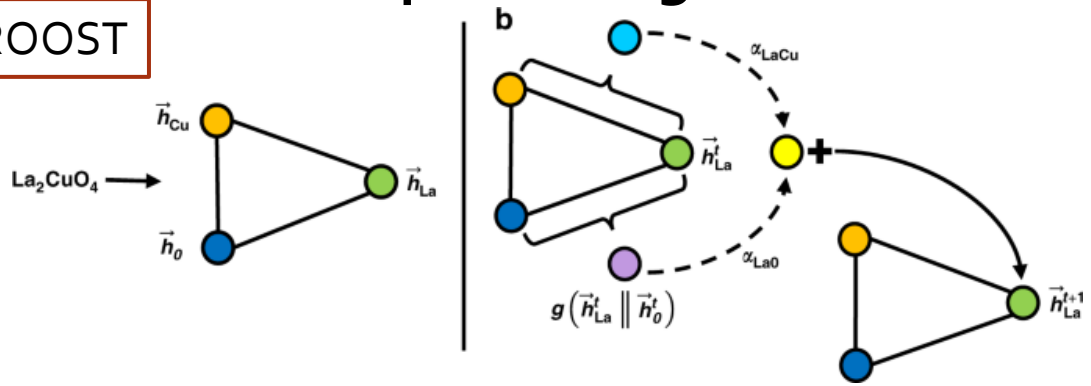
Atom2Vec



Ref: Zhou et al. PNAS. (2018)

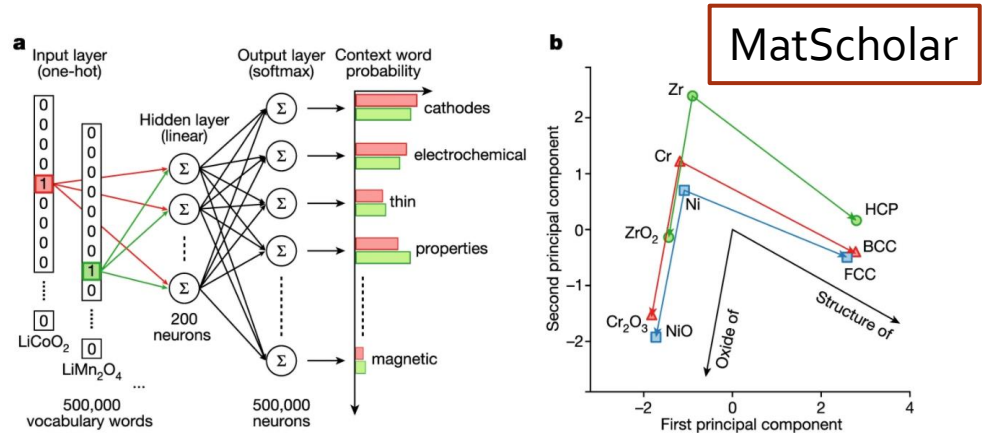
Creative Deep Learning Architectures

ROOST



Ref: Goodall and Lee. Nat Comm. (2020)

MatScholar

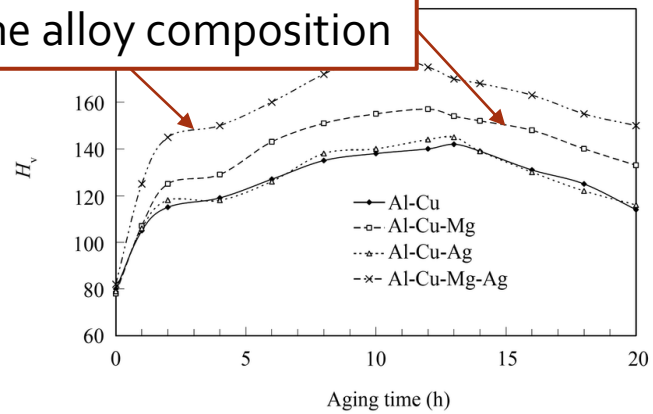


Ref: Tshitoyan et al. Nature. (2019)

Further complication: Processing and structure

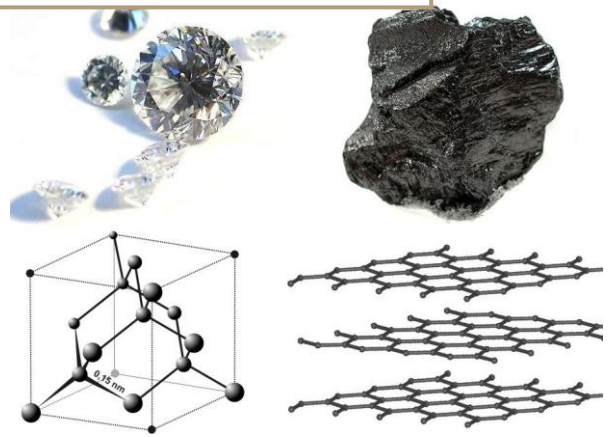
Other factors beside composition…

All same alloy composition



Ref: Song and Xiso. Sci. China. E. (2006)

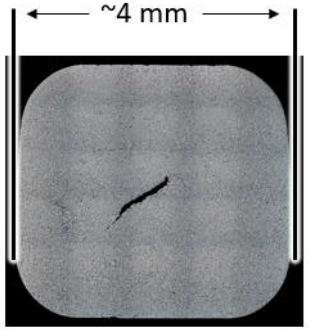
Carbon takes many forms



Ref: Wikipedia

… and you must *consider* them

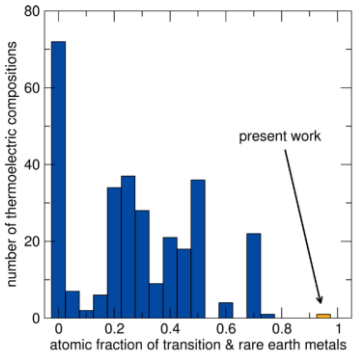
Ignoring processing or structure has given me problems in identifying materials



Ex: not accounting for casting method, alloys fractured on casting

but you can ignore and succeed

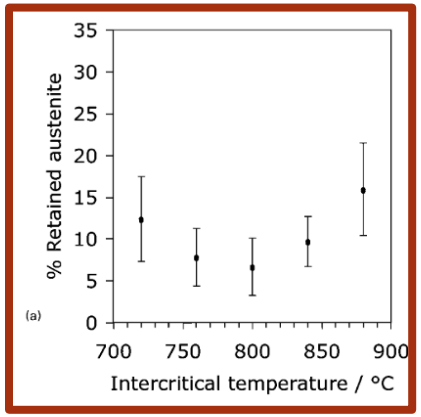
Ex: discovering new thermoelectrics based on material composition



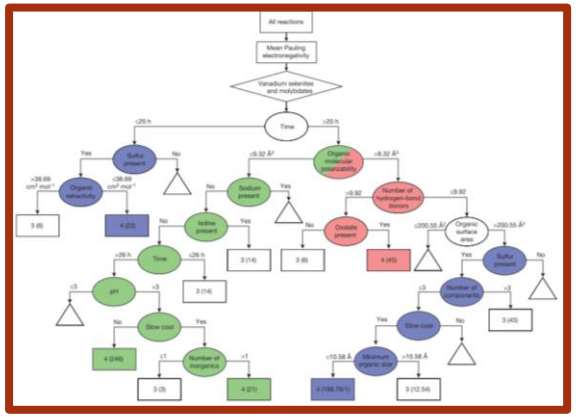
Ref: Gaultois et al. APL Materials. (2016)

There are ways for accounting for processing

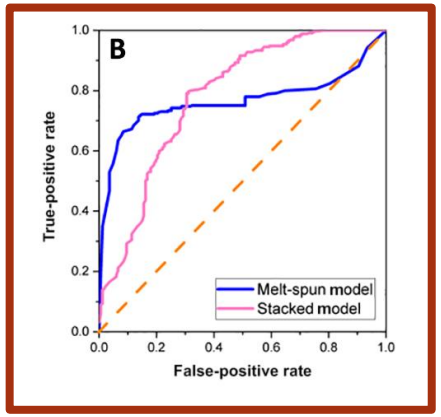
Processing parameters are easy to add to a model's inputs



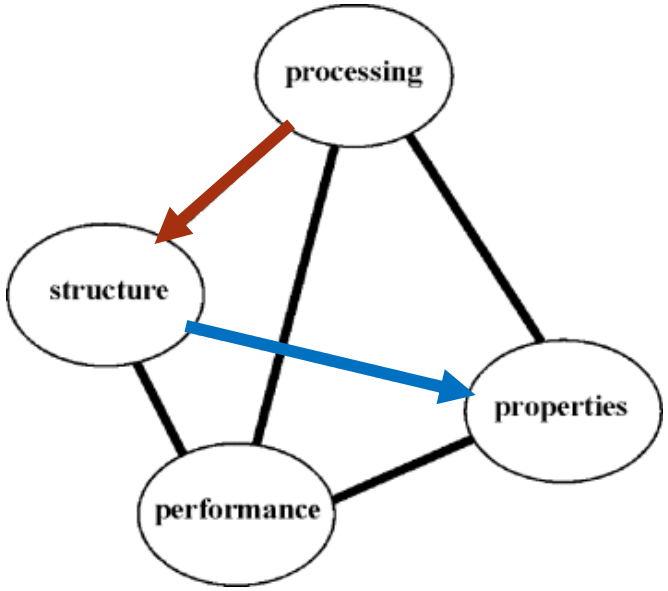
Ref: Chatterjee (2010)



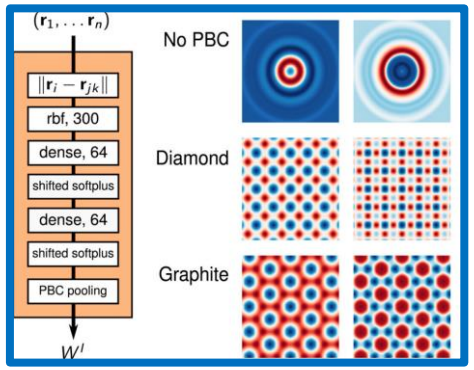
Ref: Raccuglia (2016)



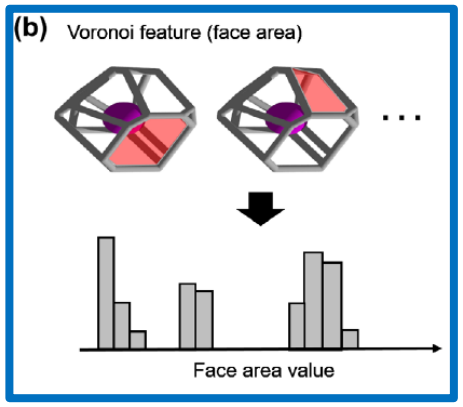
Ref: Ren (2018)



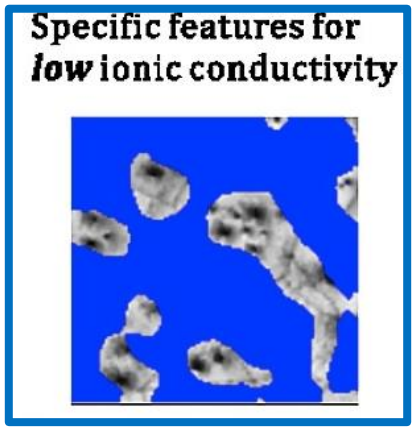
Many techniques for incorporating structure at different scales



Ref: Schütt (2018)



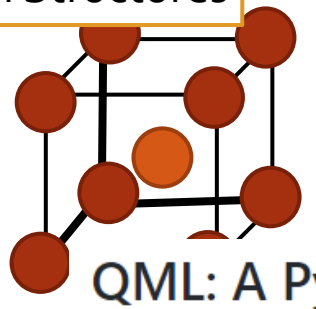
Ref: Jalem (2018)



Ref: Kondo (2017)

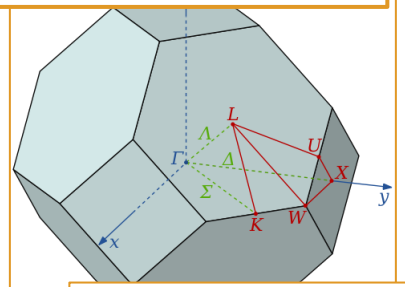
Still a new field, but already many open codes

Crystal Structures



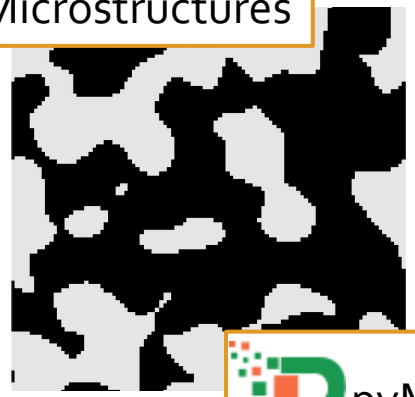
QML: A Python Toolkit

Electronic Structures



matminer

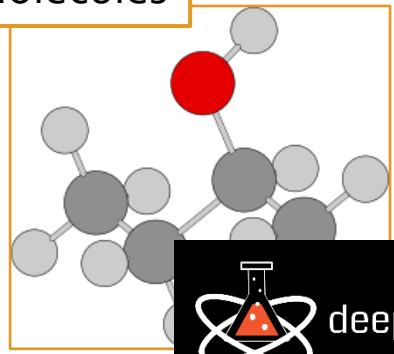
Microstructures



pyMKS

There may be a code that does what you need

Molecules

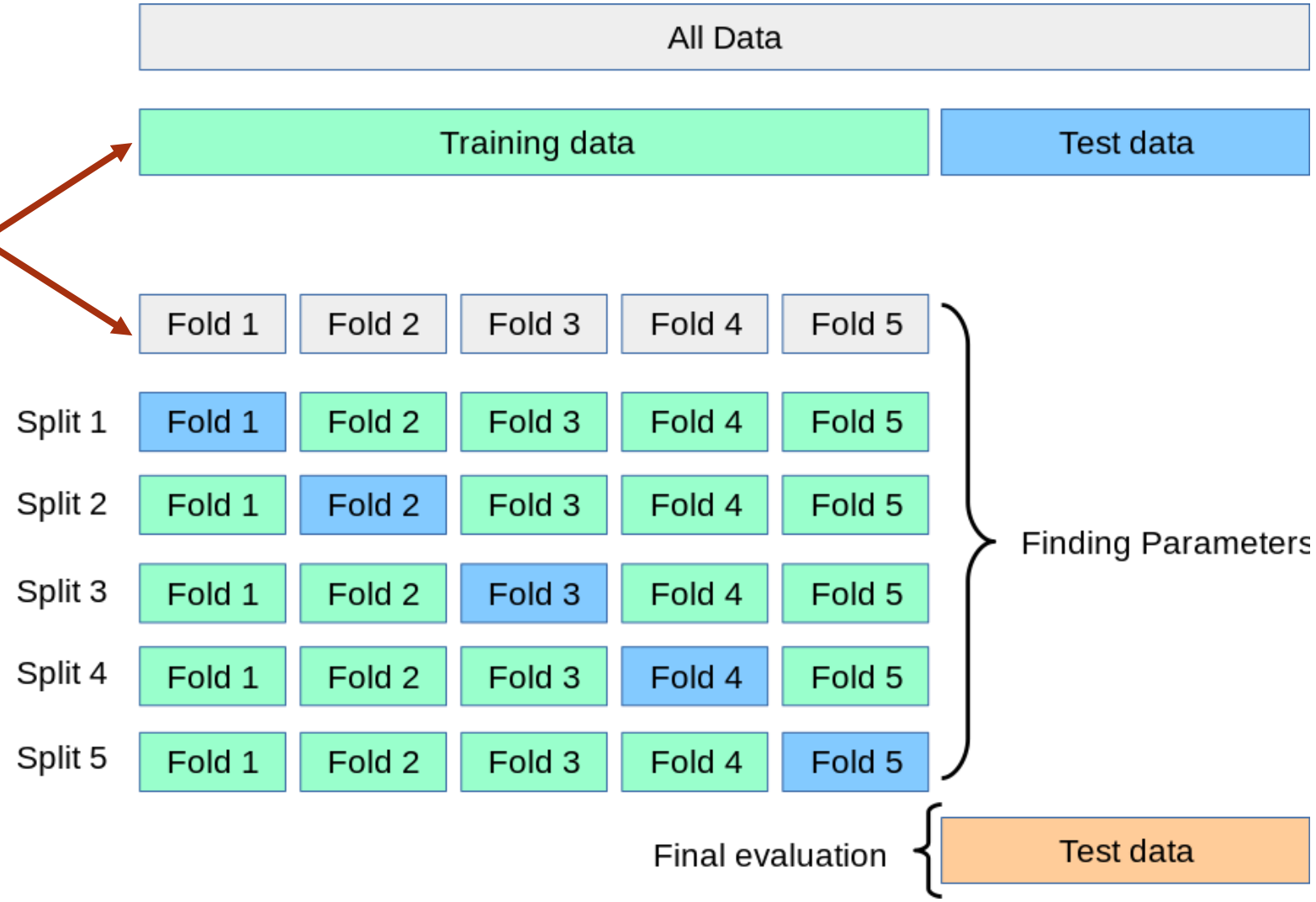


deepchem

RELATED CONTENT: TESTING YOUR MODEL ADEQUATELY

The standard approach: Random split cross-validation

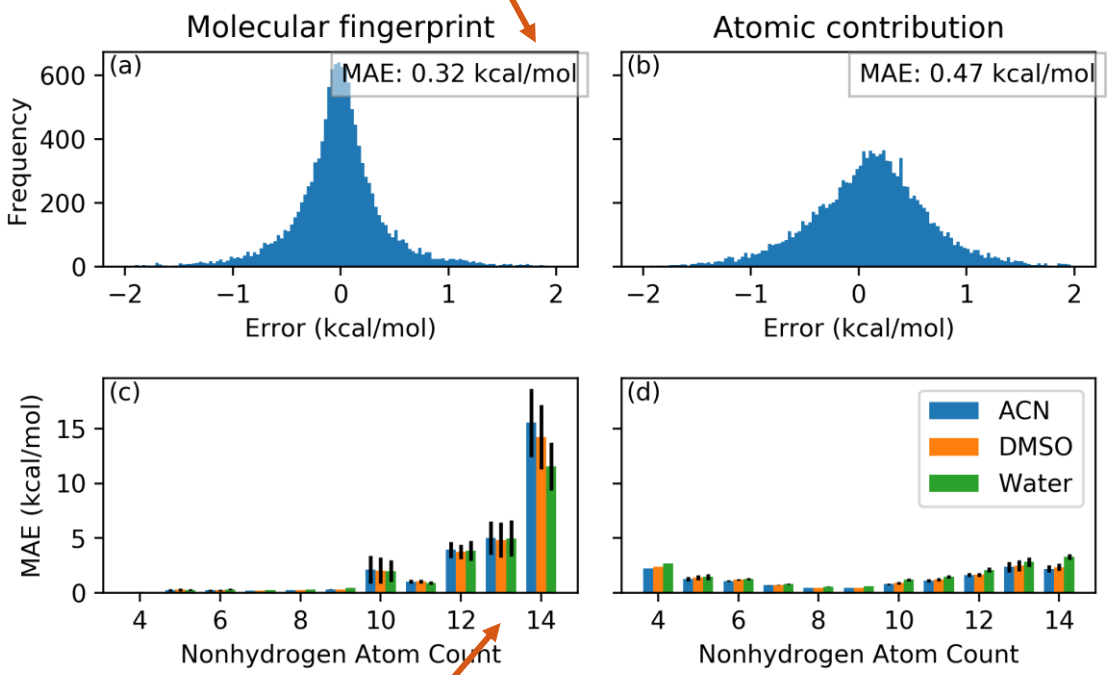
How these splits are made can lead to misleading results



Source: Scikit-learn Docs

Randomized cross-validation can be a big problem

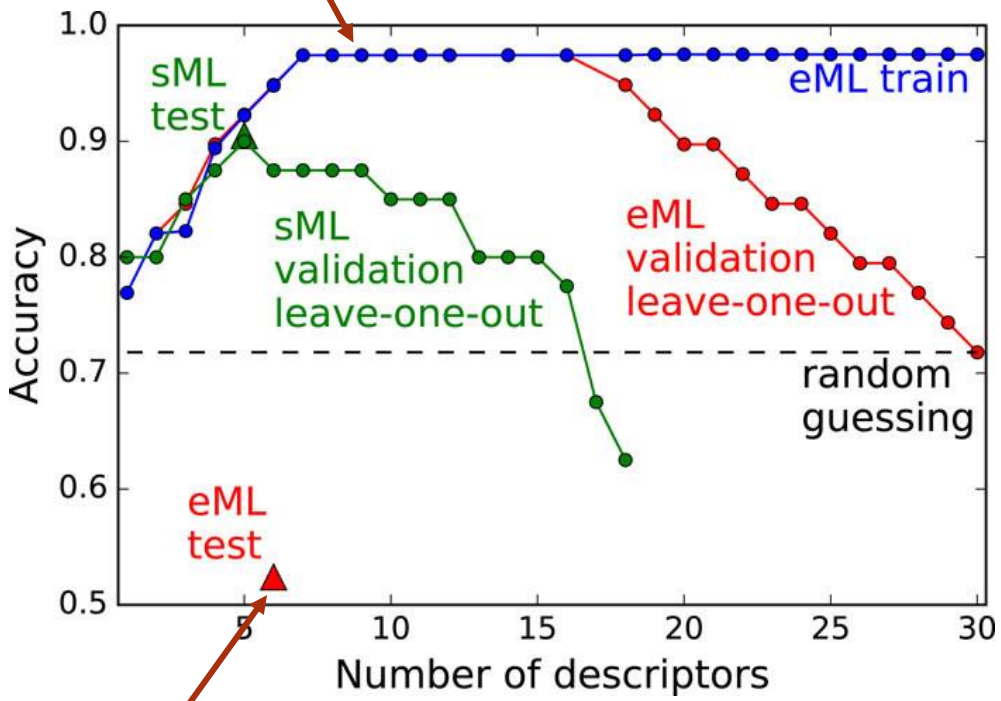
Better model on validation set



Giant errors on test set

Source: Ward et al, *in preparation*

Better model on validation set



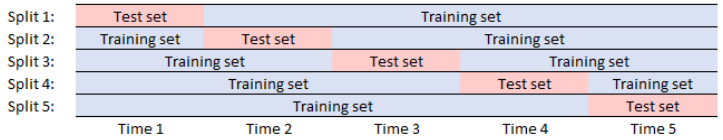
Giant errors on test set

Source: Cubuk et al. JCP. (2019)

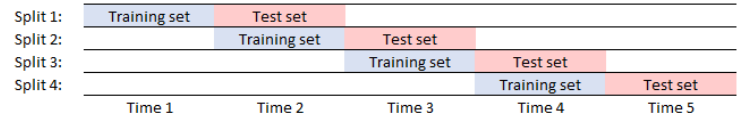
Designing validation to mimic practice

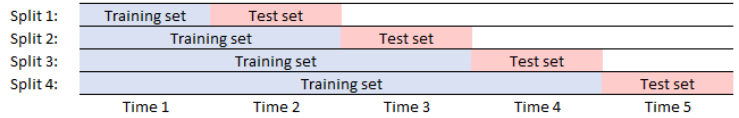
Time Series

✗ Predicting past *and* future



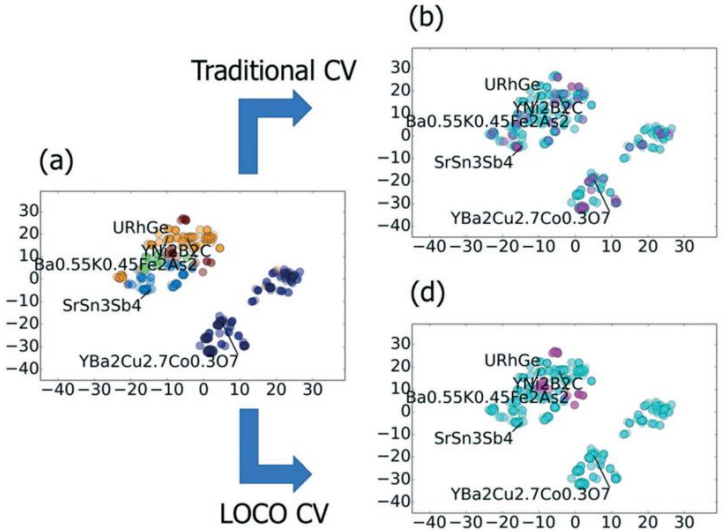
✔ Predict only the future





Clustered data

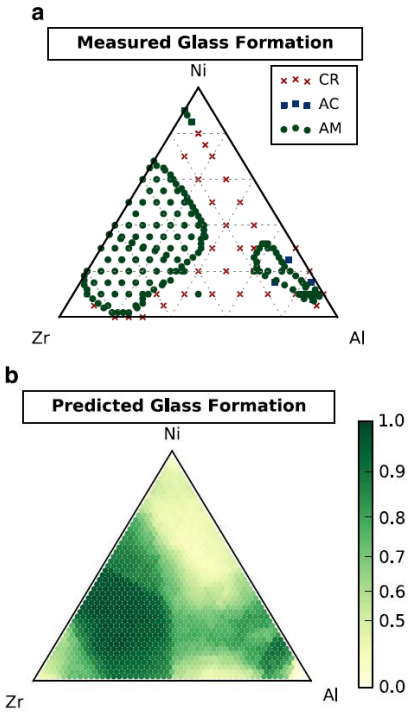
✗ Ignoring clustering in data



✔ Leave out by cluster

Alloy materials

✔ Excluding an alloy system



Source: CrossValidated

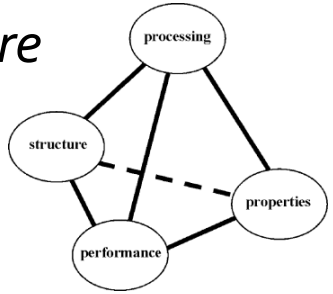
Ref: Meredig et al. MSDE (2018)

Ref: Ward et al. npj Comp Mat. (2016)

Wrap up and take-home points

Main Lecture: Machine learning for materials

- Materials have *processing* and *structure*



- Descriptors for inorganics are similar:
  - Elemental-property-based features for compositions
  - Account for structure and properties <u>if needed</u> (e.g., n-point correlation features)

- Many codes exist!

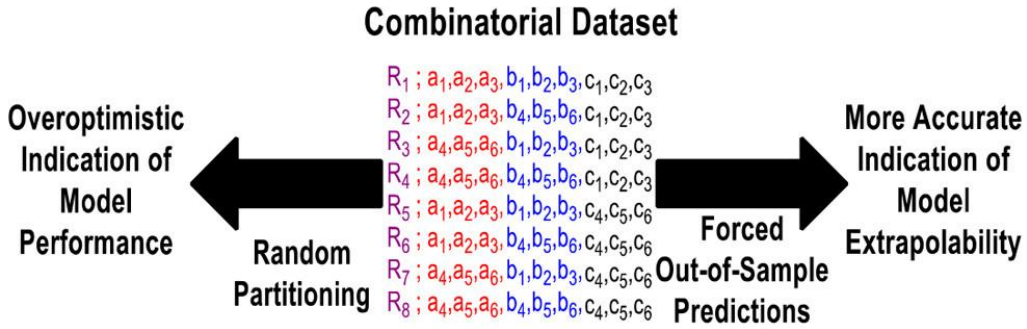
QML: A Python Toolkit

pyMKS

matminer

Side note: Validate like you mean to use it

<u>Scientific data is often biased.</u>



Ref: Zahrt et al. ACS Combi. (2020)

<u>Consider that bias in your validation</u>

Practical assignments on GitHub will illustrate both concepts!