Summary of key points (so far)

CEMS
Chemical Engineering
& Materials Science

1) General supervised learning:

2) Success depends on:

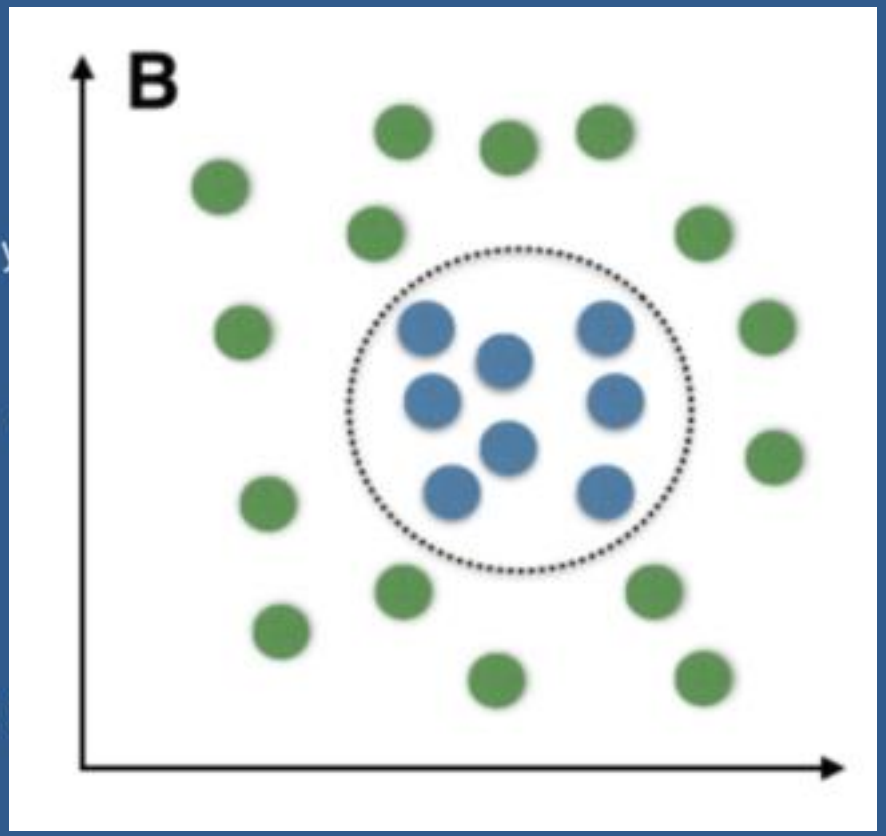
3) Validation is essential:

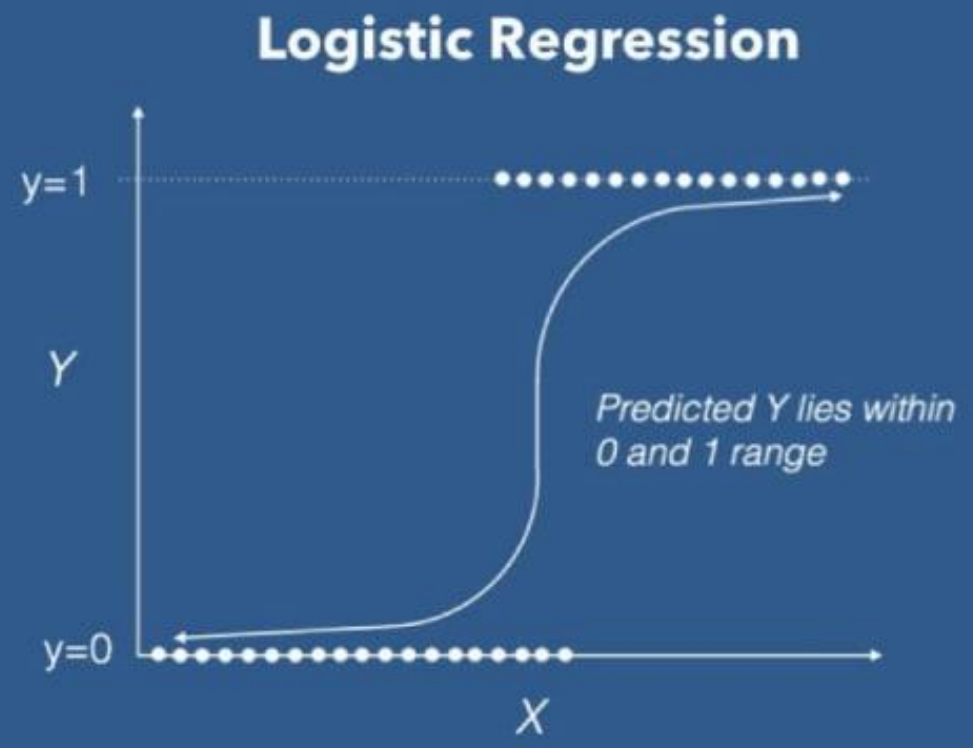
4) Leverage your domain knowledge:

What about nonlinear classification?

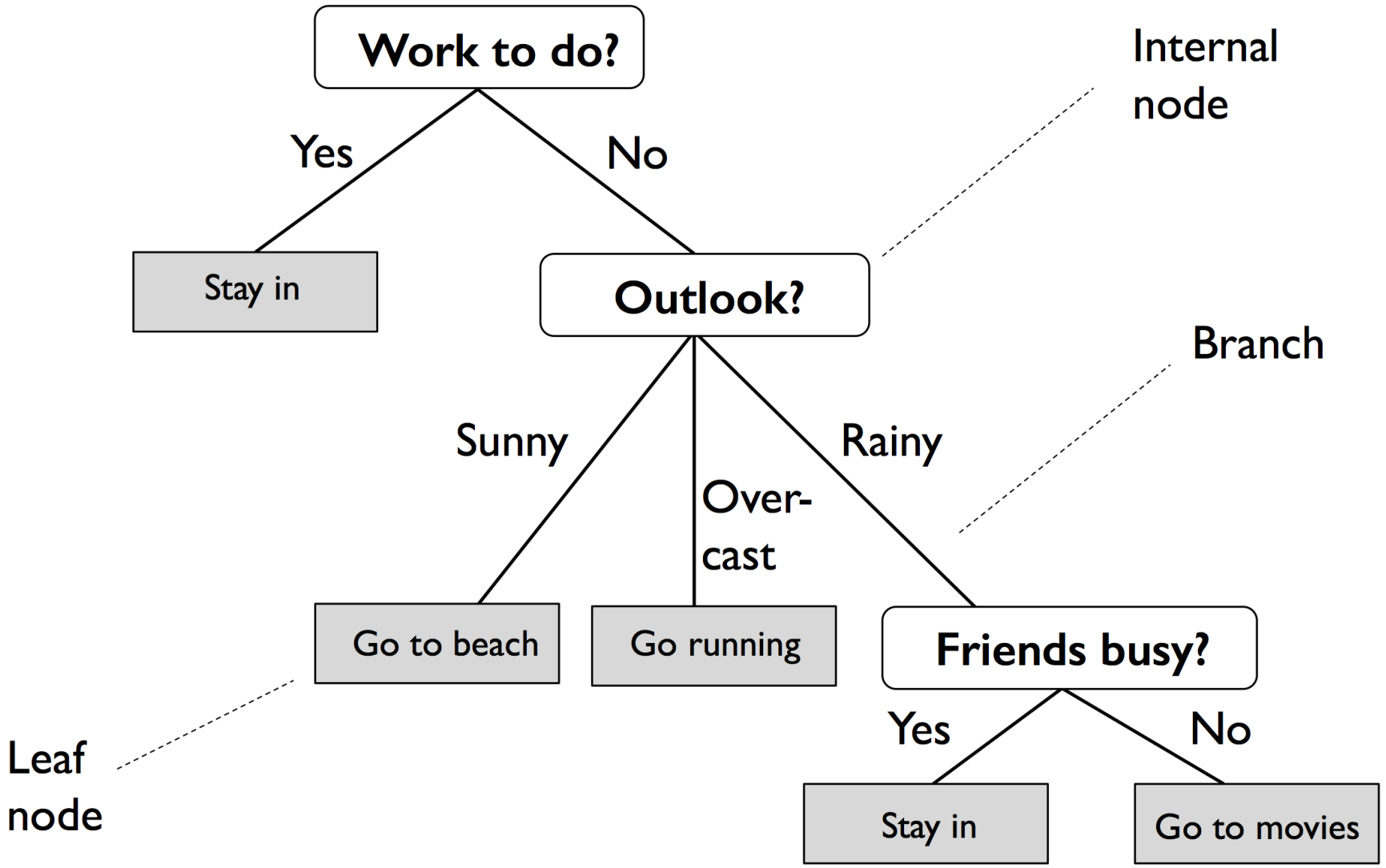
For binary classification, y is no longer continuous, but binomial:

y = [1, 1, 1, -1, -1, 1, -1, -1, ...]

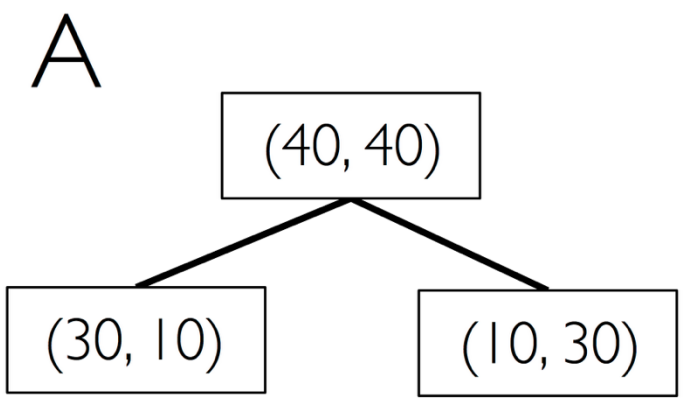


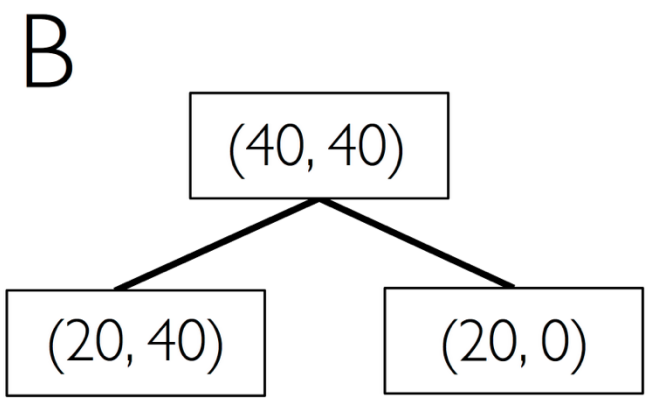


Brief intro to decision trees



Which split is better?





Brief intro to decision trees

Determine splits by **maximizing information gain (IG)**

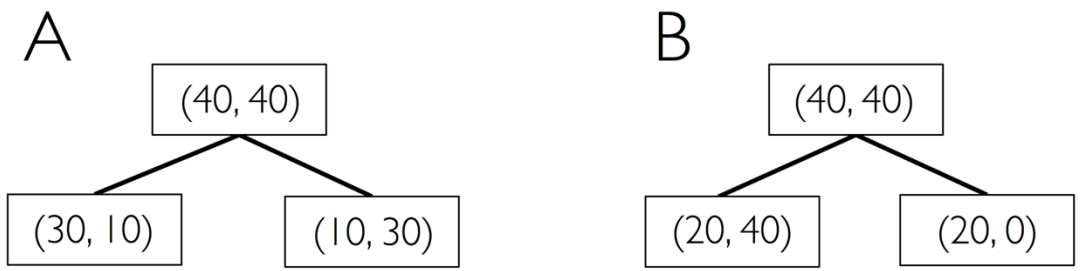
minimizing weighted impurity, I

Brief intro to decision trees

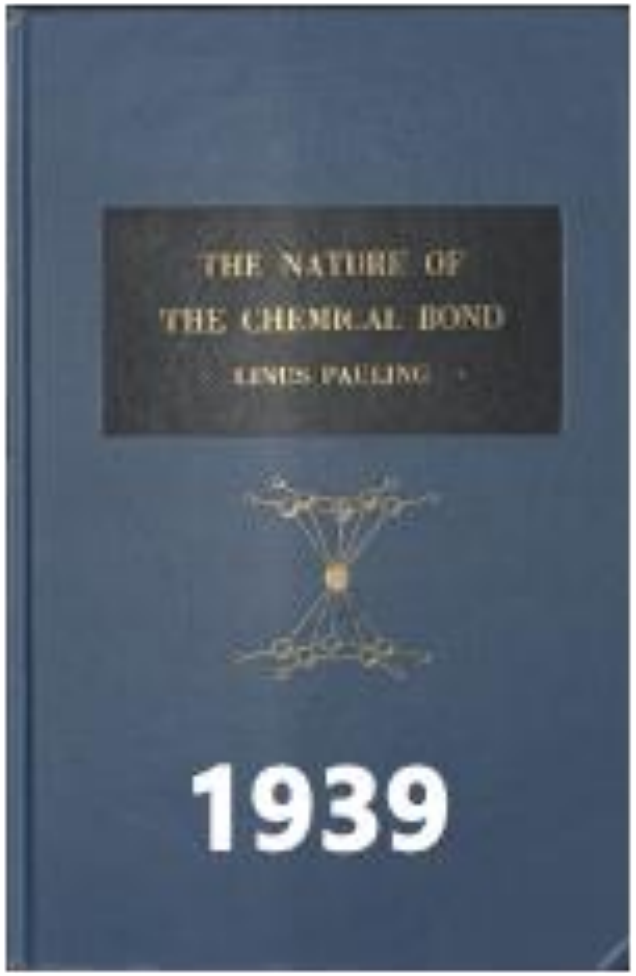
Determine splits by **maximizing information gain (IG)**

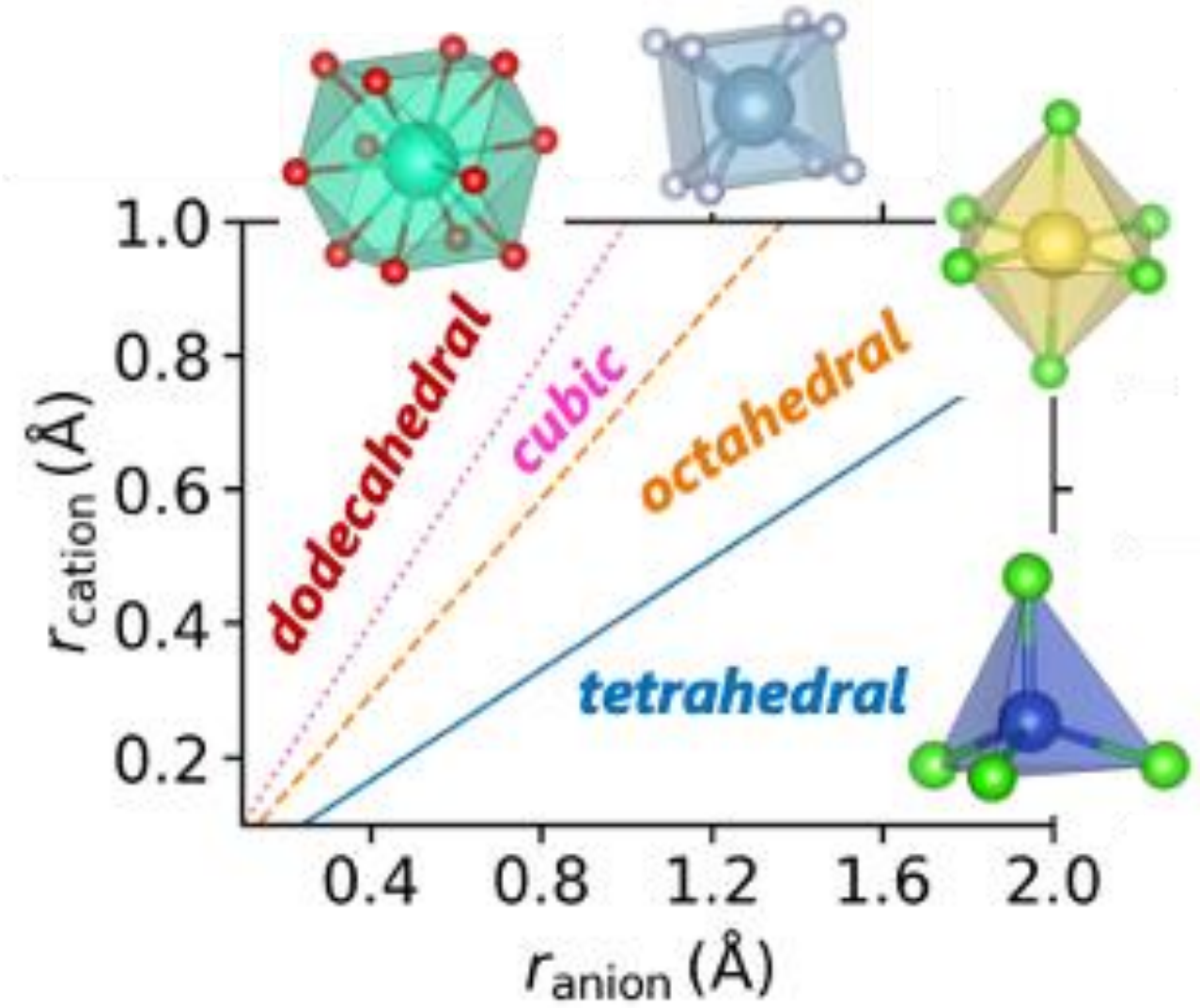
minimizing weighted impurity, I

Brief intro to decision trees

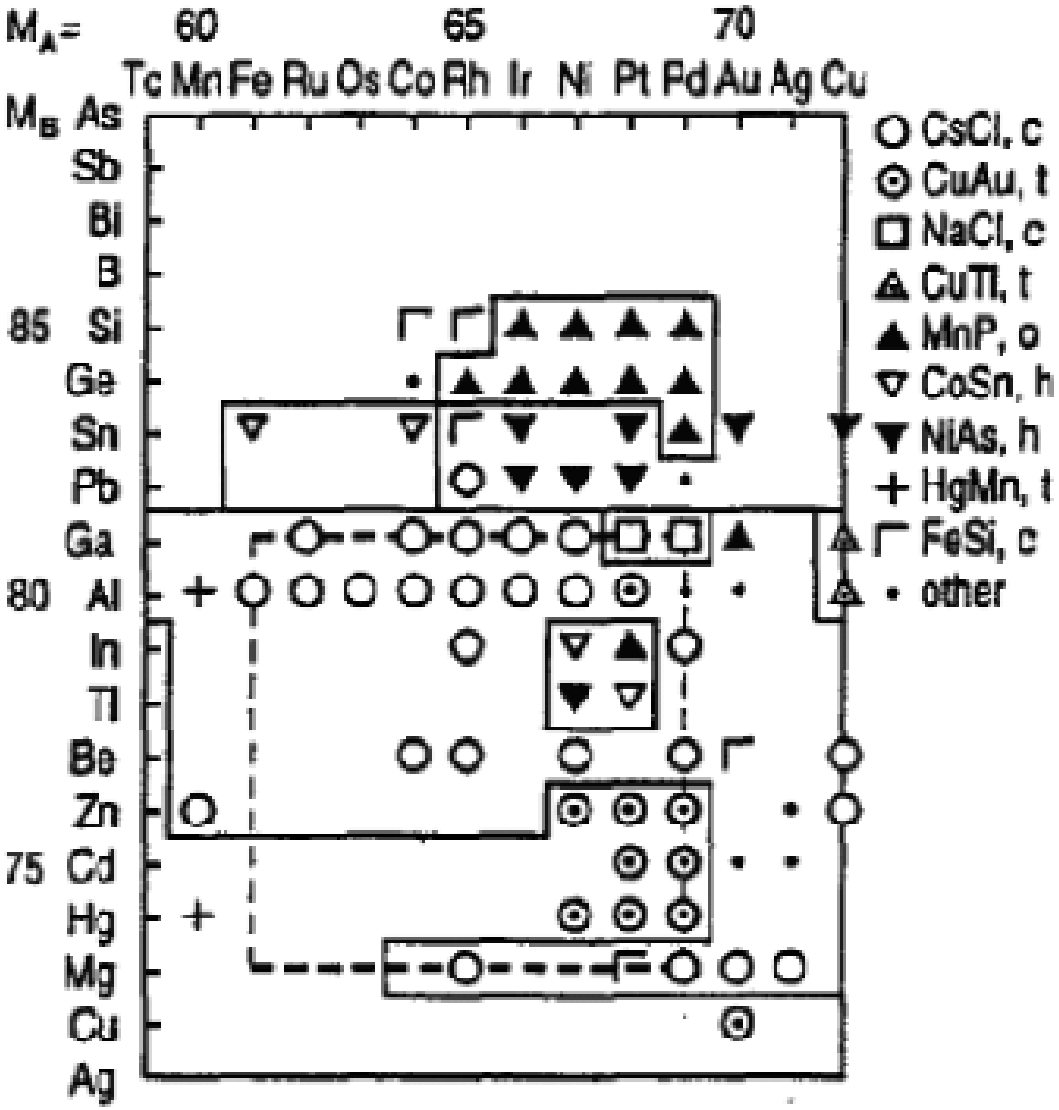


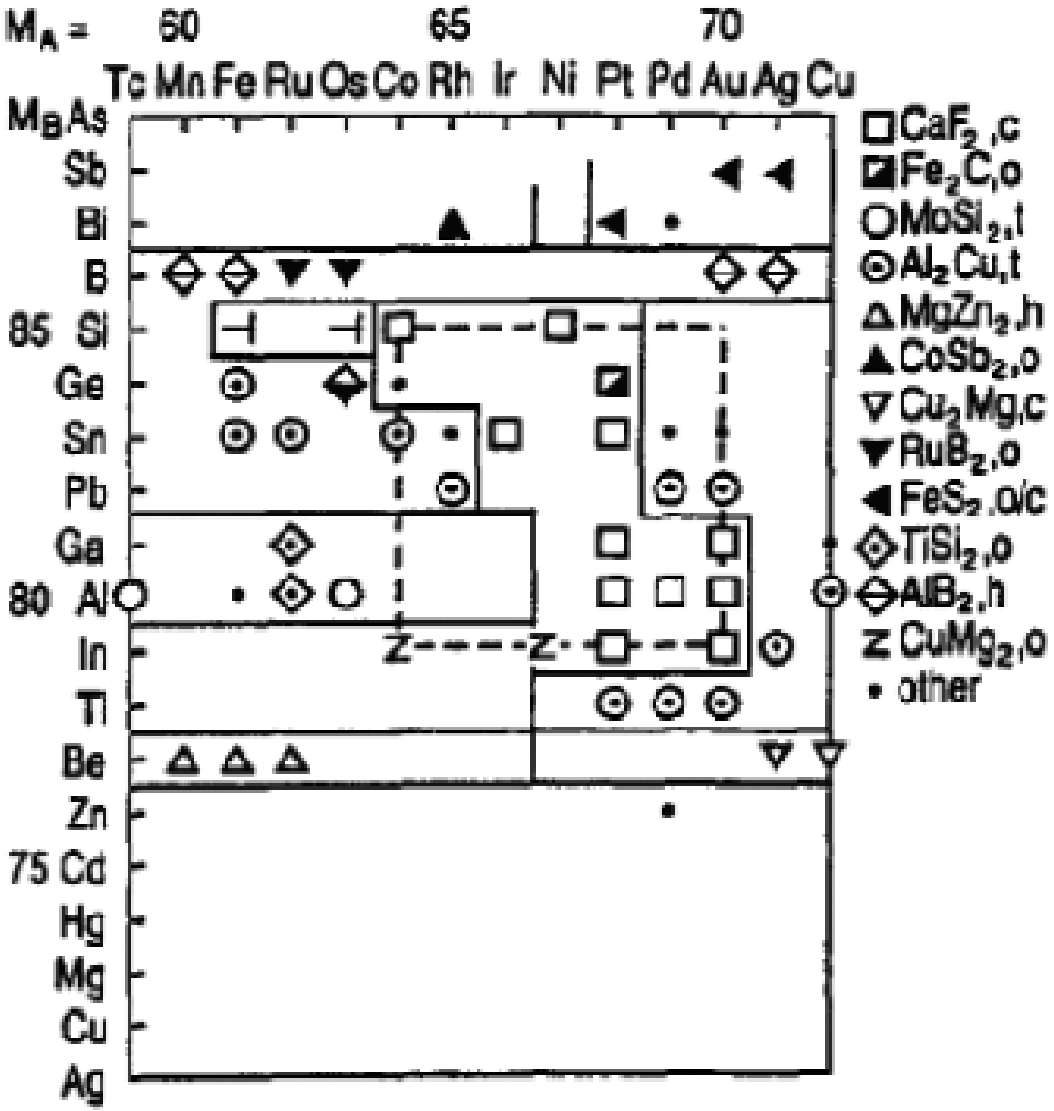
Classifying crystal structures





Classifying crystal structures



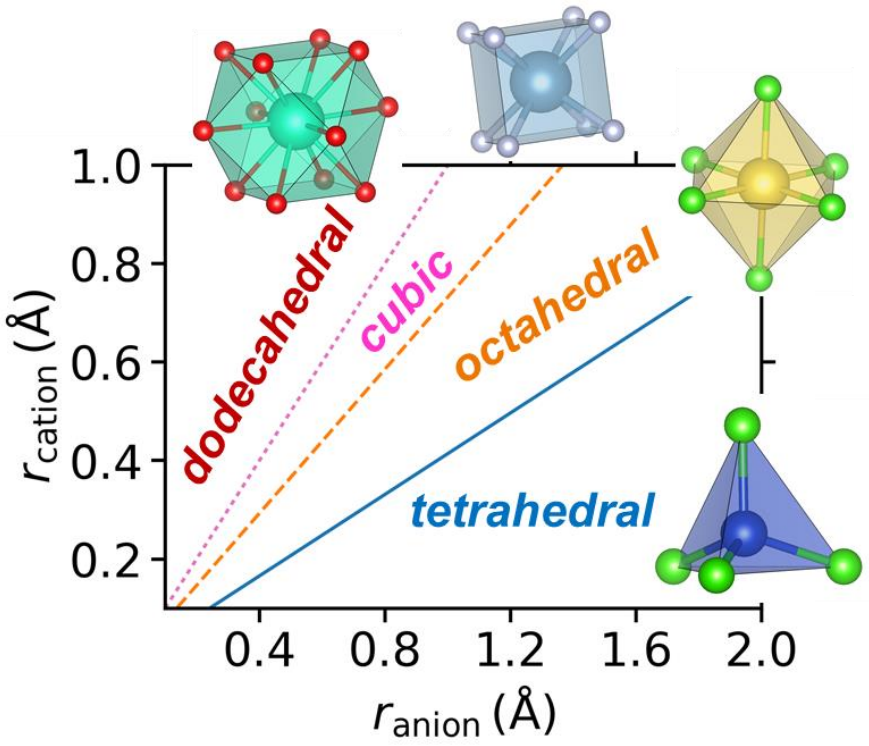


Pettifor, 1980

Finding simple models w/ supervised ML

y – *target property (observable)*

y *– data you find or generate*

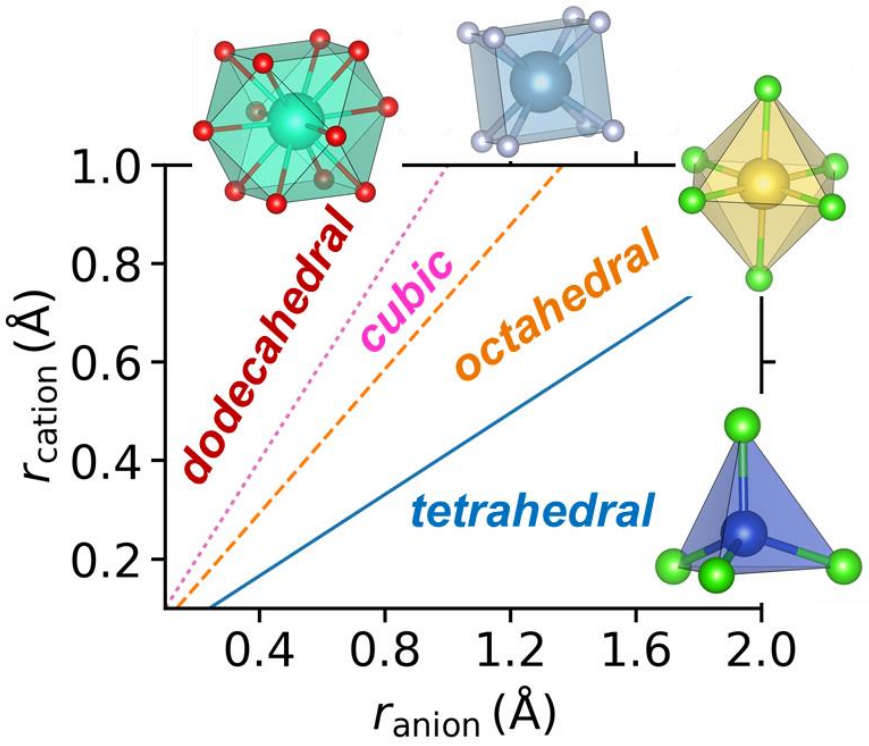


$$\mathbf{y} = \begin{bmatrix} \\ \\ \\ \\ \\ \\ \\ \\ \end{bmatrix}$$

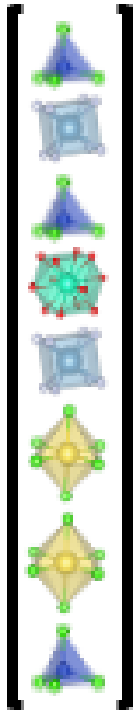
Finding simple models w/ supervised ML

y – *target property (observable)*
X **–** *feature space (representation)*

y *– data you find or generate*
X *– stuff you hope relates to* ***y***



$$\mathbf{y} \qquad \boldsymbol{X}$$



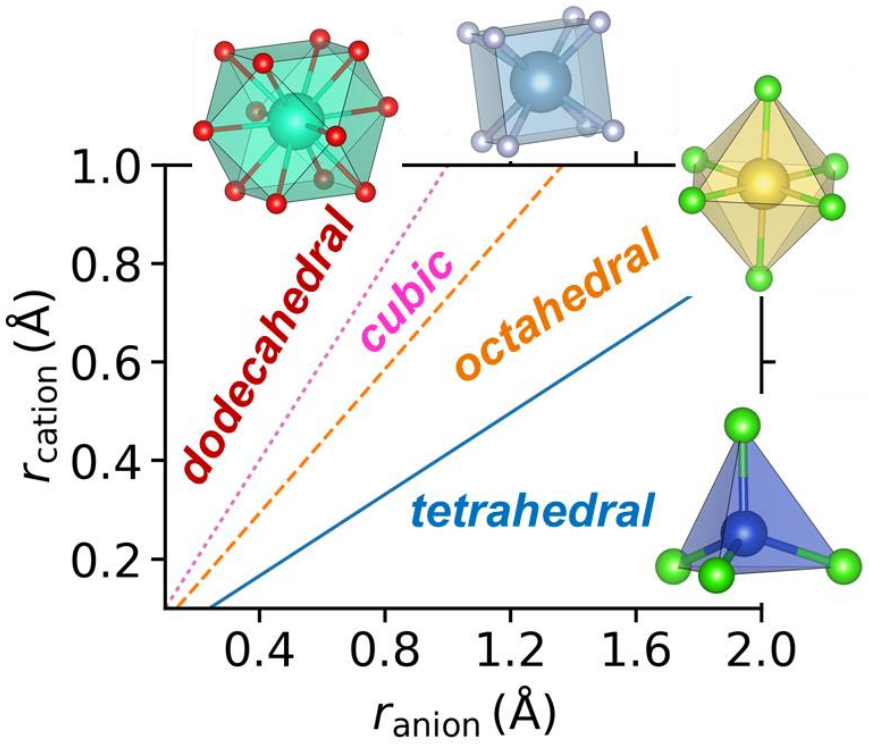
$$\begin{bmatrix} r_{cation} & r_{anion} \\ 0.5 & 1.7 \\ 0.7 & 1.1 \\ 0.3 & 1.2 \\ \ldots & \ldots \end{bmatrix}$$

Finding simple models w/ supervised ML

CEMS
Chemical Engineering & Materials Science

y – *target property (observable)*
X – *feature space (representation)*
***f*(*X*)** – *model (descriptor)*
ŷ – *prediction (model output)*

y – *data you find or generate*
X – *stuff you hope relates to* ***y***
f – *the learned mapping of* ***X*** *to* ***y***



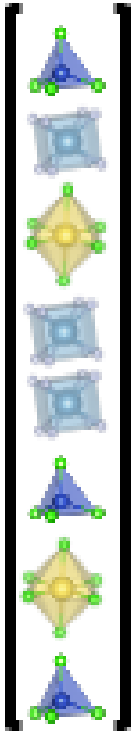
y

X

r_{cation}	r_{anion}
0.5	1.7
0.7	1.1
0.3	1.2
…	…

***f*(*X*): min(y-ŷ)**

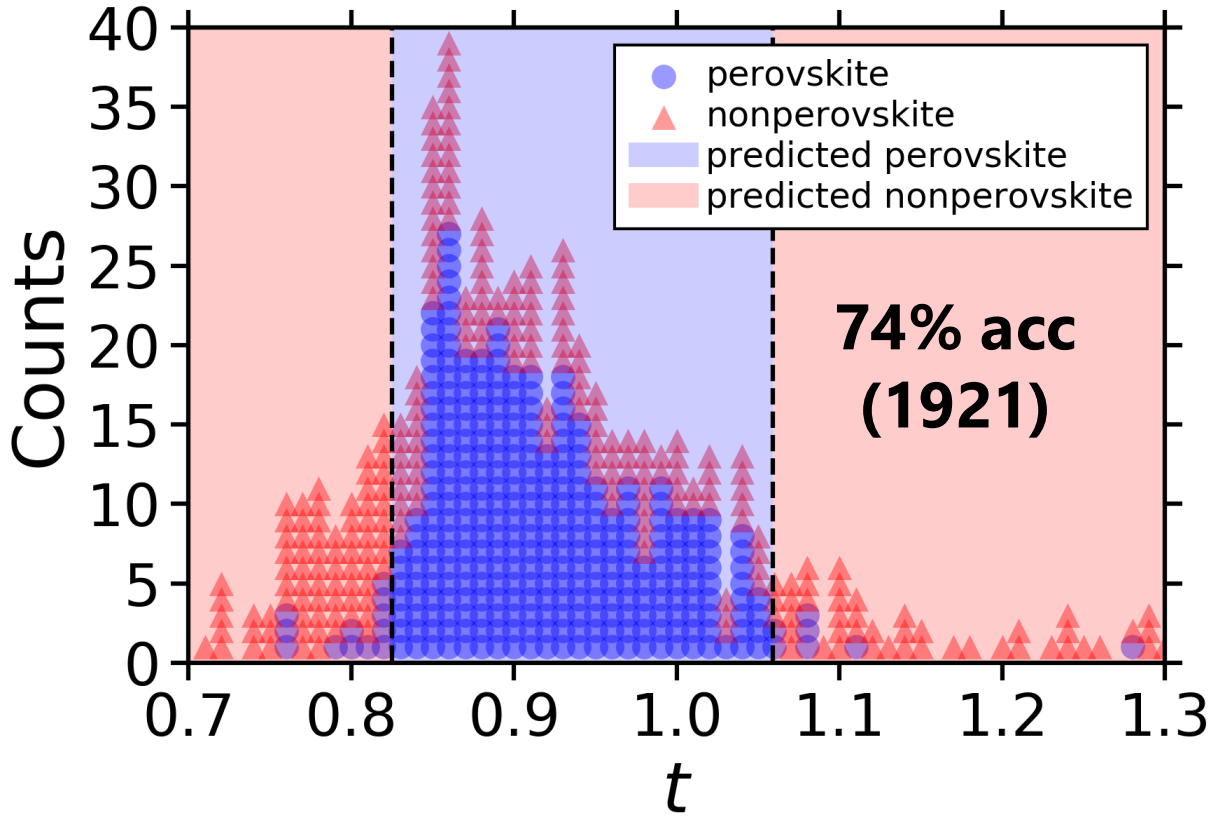
ŷ

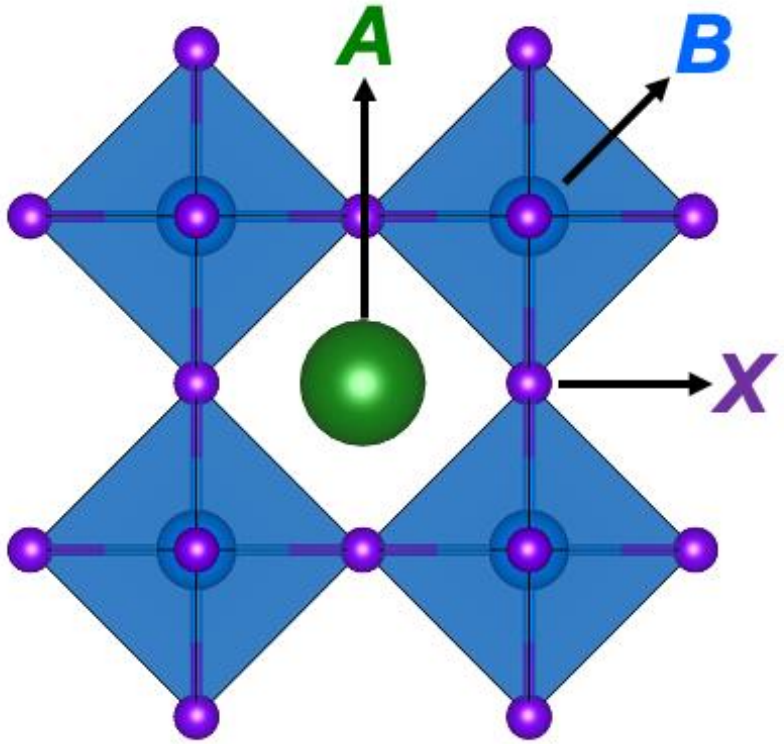


Goldschmidt's tolerance factor for perovskite stability

For 576 experimentally characterized ABX_3 compounds

$$t = \frac{r_A + r_X}{\sqrt{2}(r_B + r_X)}$$





Bartel et al., *Science Advances*, 2019, 5, eaav0693

New tolerance factor!

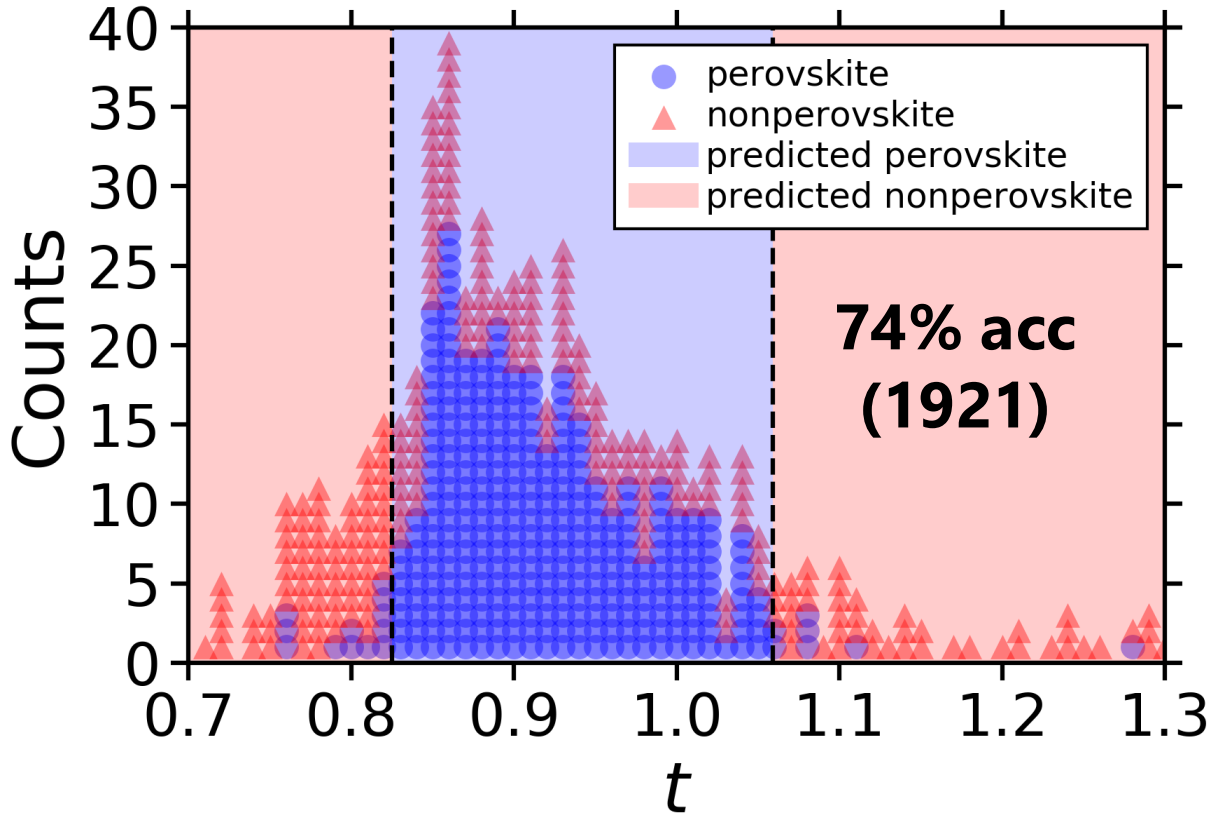
CEMS
Chemical Engineering & Materials Science

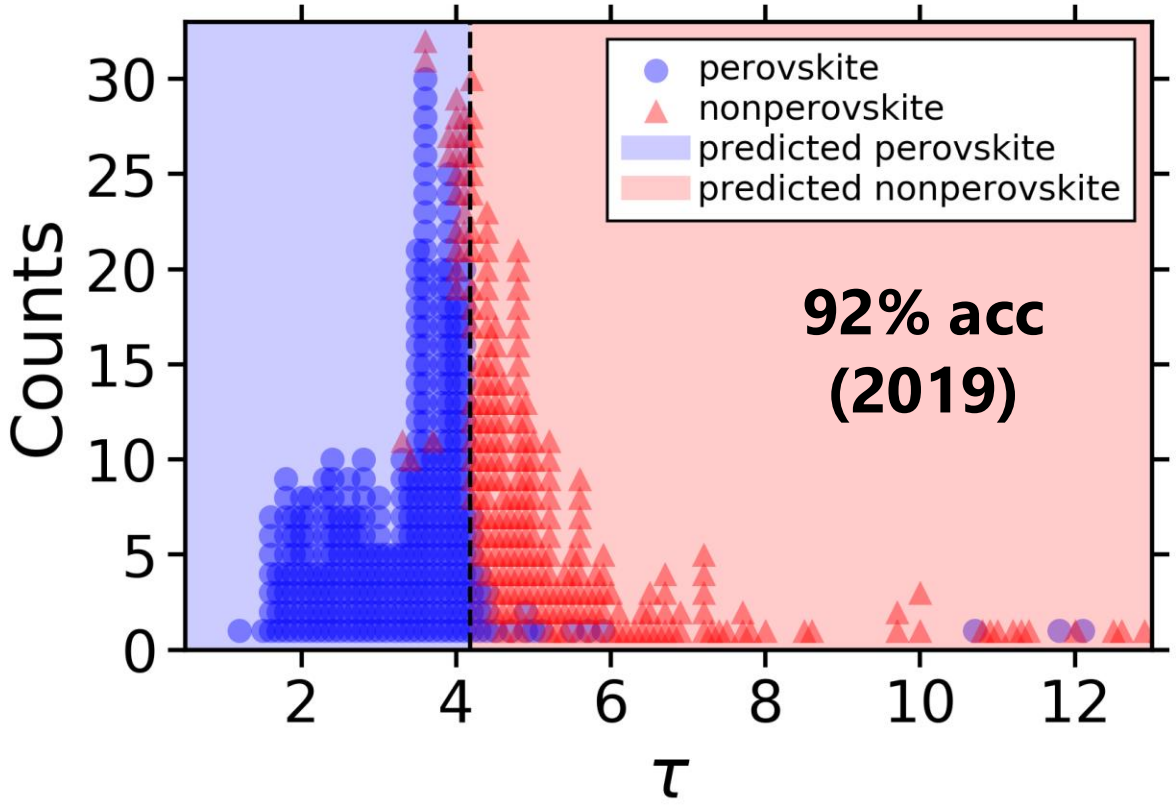
For 576 experimentally characterized ABX_3 compounds

$$t = \frac{r_A + r_X}{\sqrt{2}(r_B + r_X)}$$

SISSO

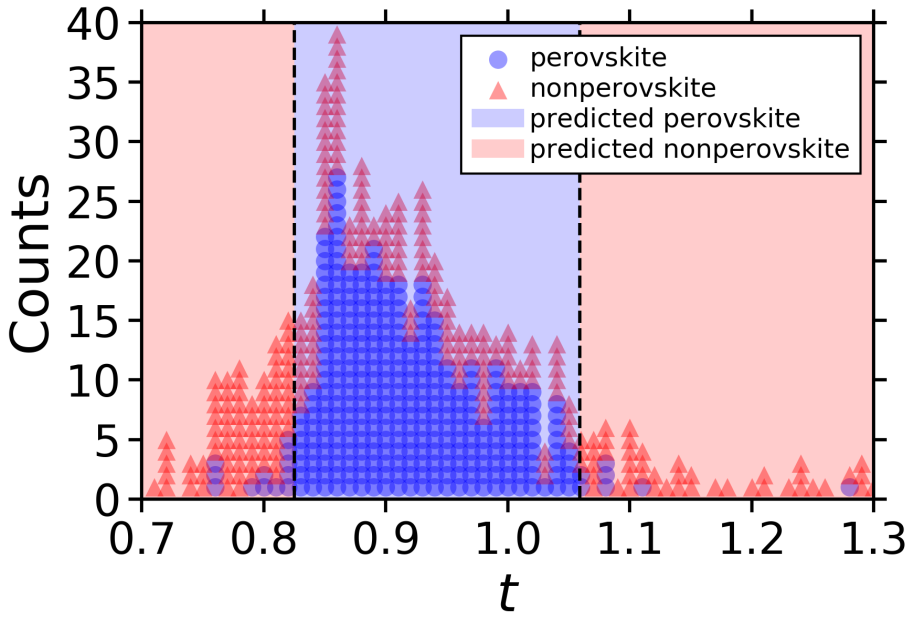
$$\tau = \frac{r_X}{r_B} - n_A\left(n_A - \frac{r_A/r_B}{\ln r_A/r_B}\right)$$

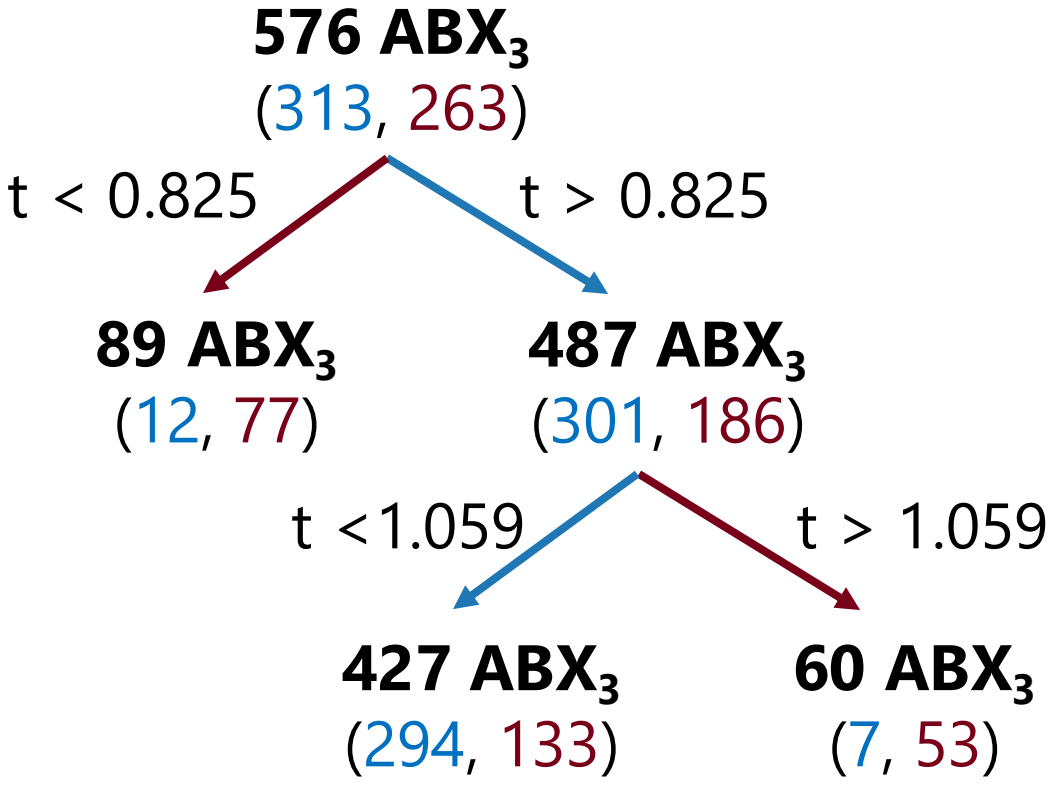




Bartel et al., *Science Advances*, 2019, 5, eaav0693

Decision trees w/ Goldschmidt's t





y

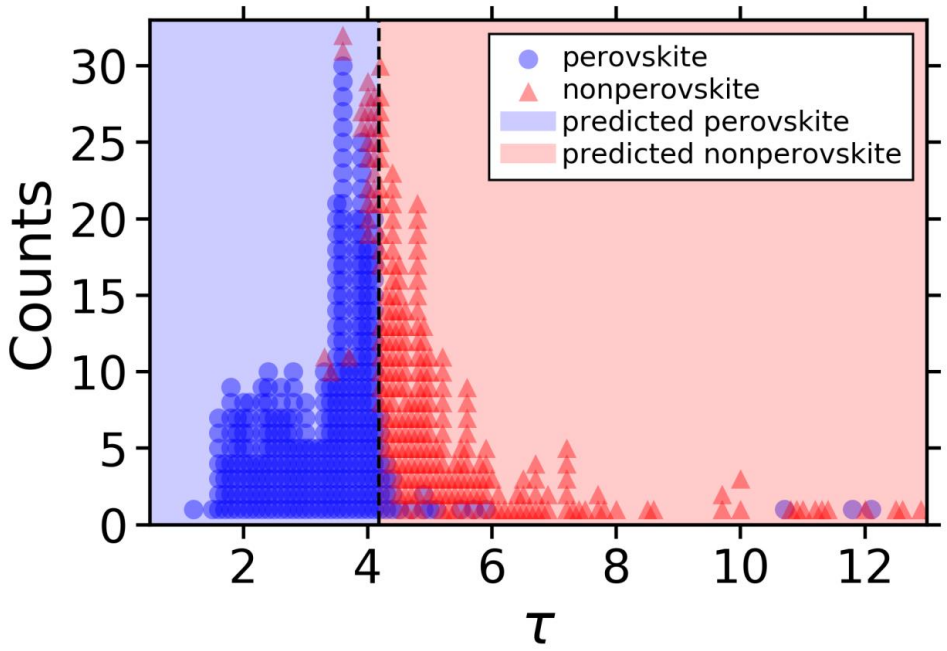
$$\begin{bmatrix} 1 \\ -1 \\ 1 \\ \dots \end{bmatrix}$$

X

$$\begin{bmatrix} t = \dfrac{r_A + r_X}{\sqrt{2}(r_B + r_X)} \end{bmatrix}$$

Bartel et al., *Science Advances*, 2019, 5, eaav0693

Decision trees w/ τ



576 ABX_3
(313, 263)

$\tau < 4.18$ → **321 ABX_3** (293, 28)

$\tau > 4.18$ → **255 ABX_3** (20, 235)

y

$$\begin{bmatrix} 1 \\ -1 \\ 1 \\ \dots \end{bmatrix}$$

X

$$\begin{bmatrix} \tau = \frac{r_X}{r_B} - n_A\left(n_A - \frac{r_A/r_B}{\ln r_A/r_B}\right) \\ \\ \end{bmatrix}$$

Bartel et al., *Science Advances*, 2019, 5, eaav0693