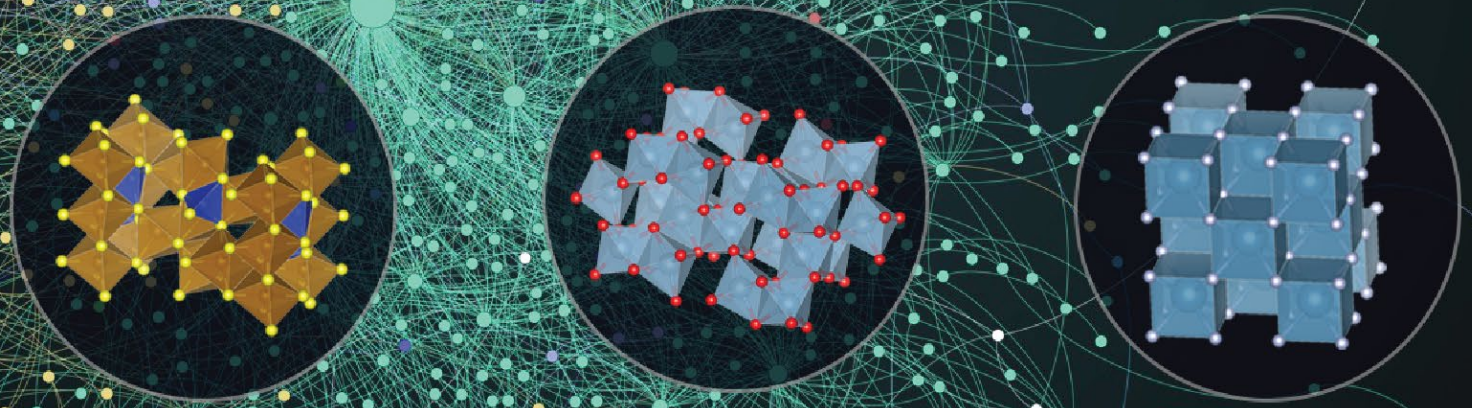
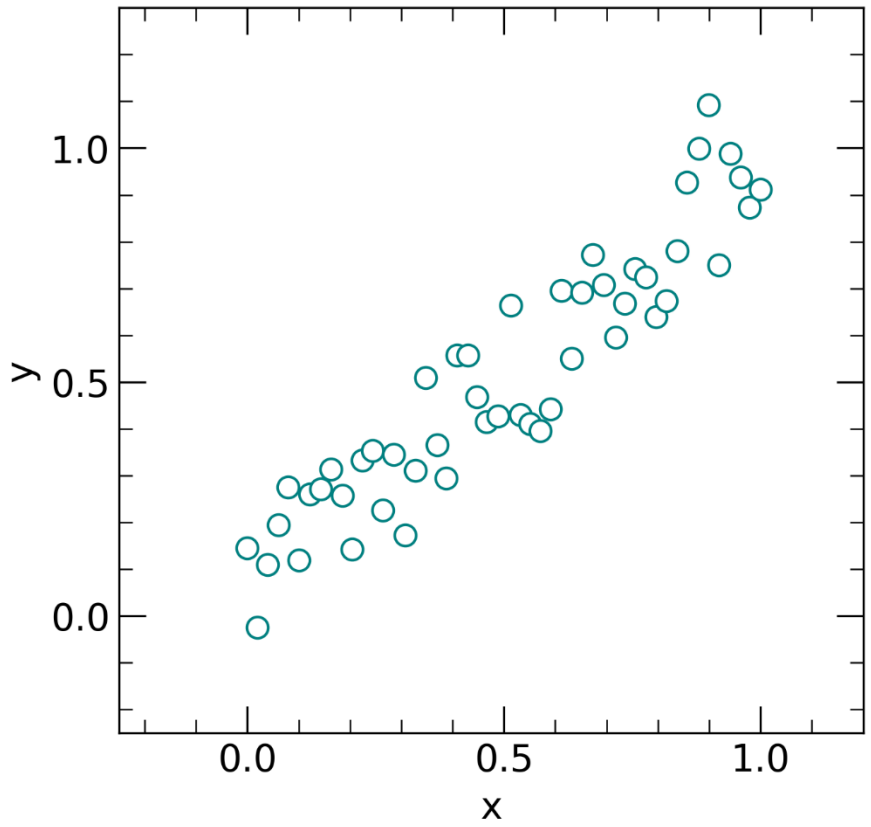
Extrapolation in Materials Informatics

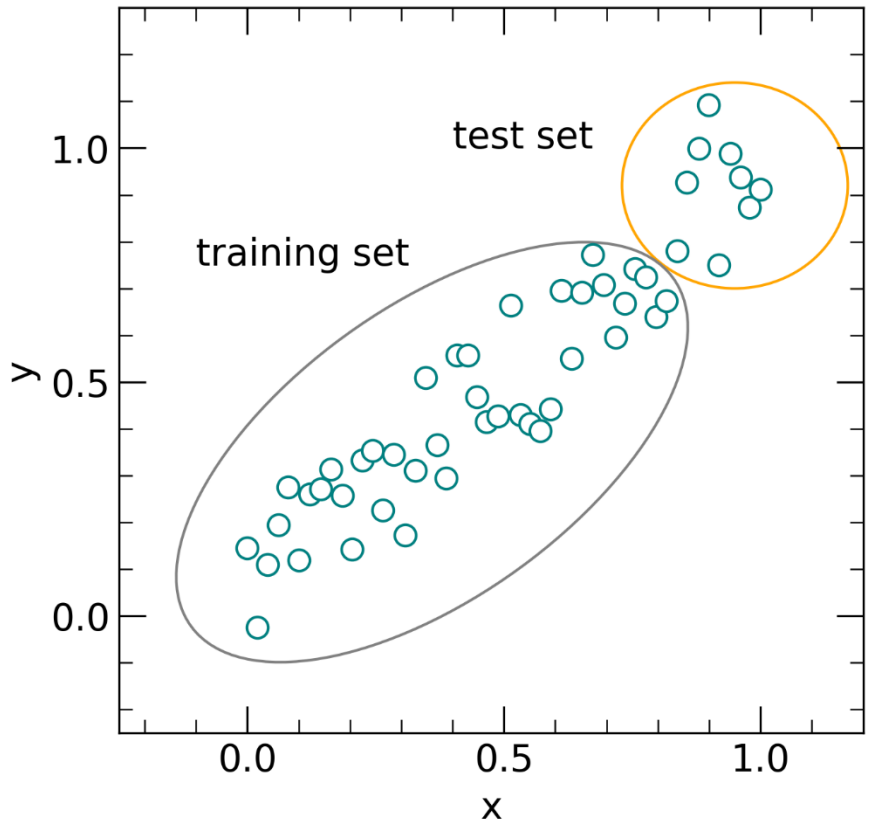


Most machine learning approaches fail with extrapolation

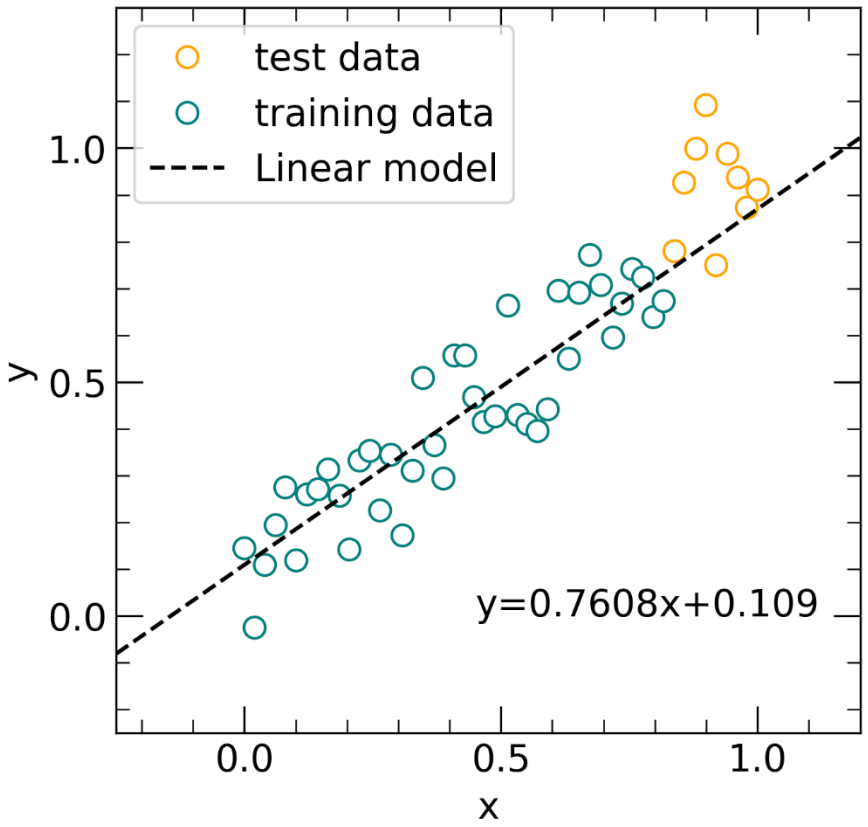


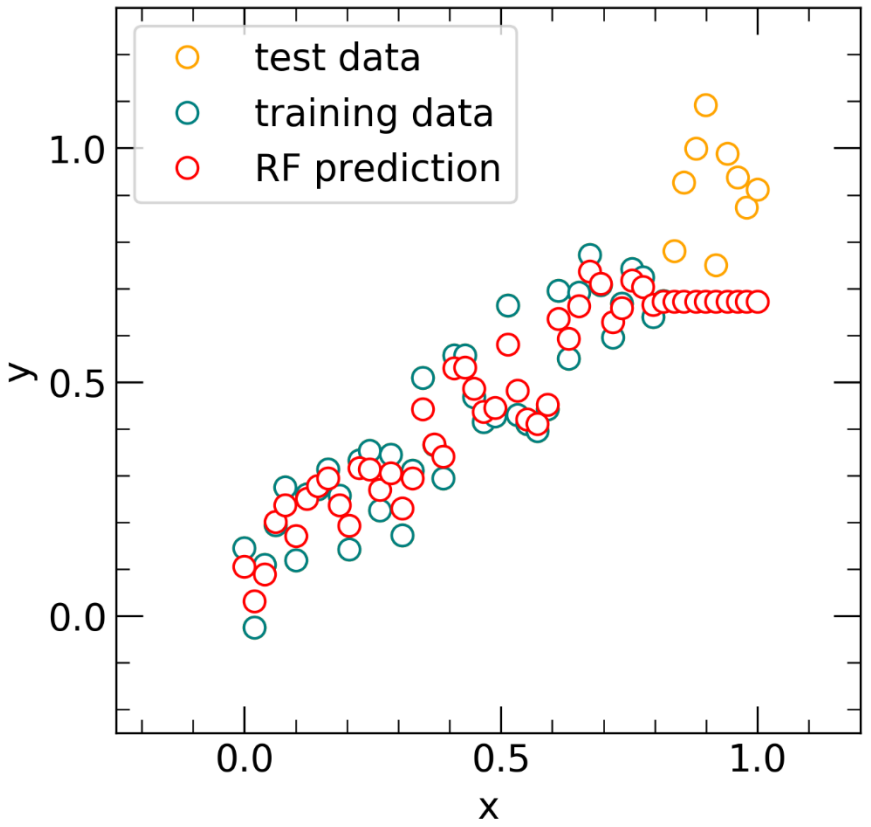
Most machine learning approaches fail with extrapolation

Will we be able to predict the validation set if these are extrapolations?



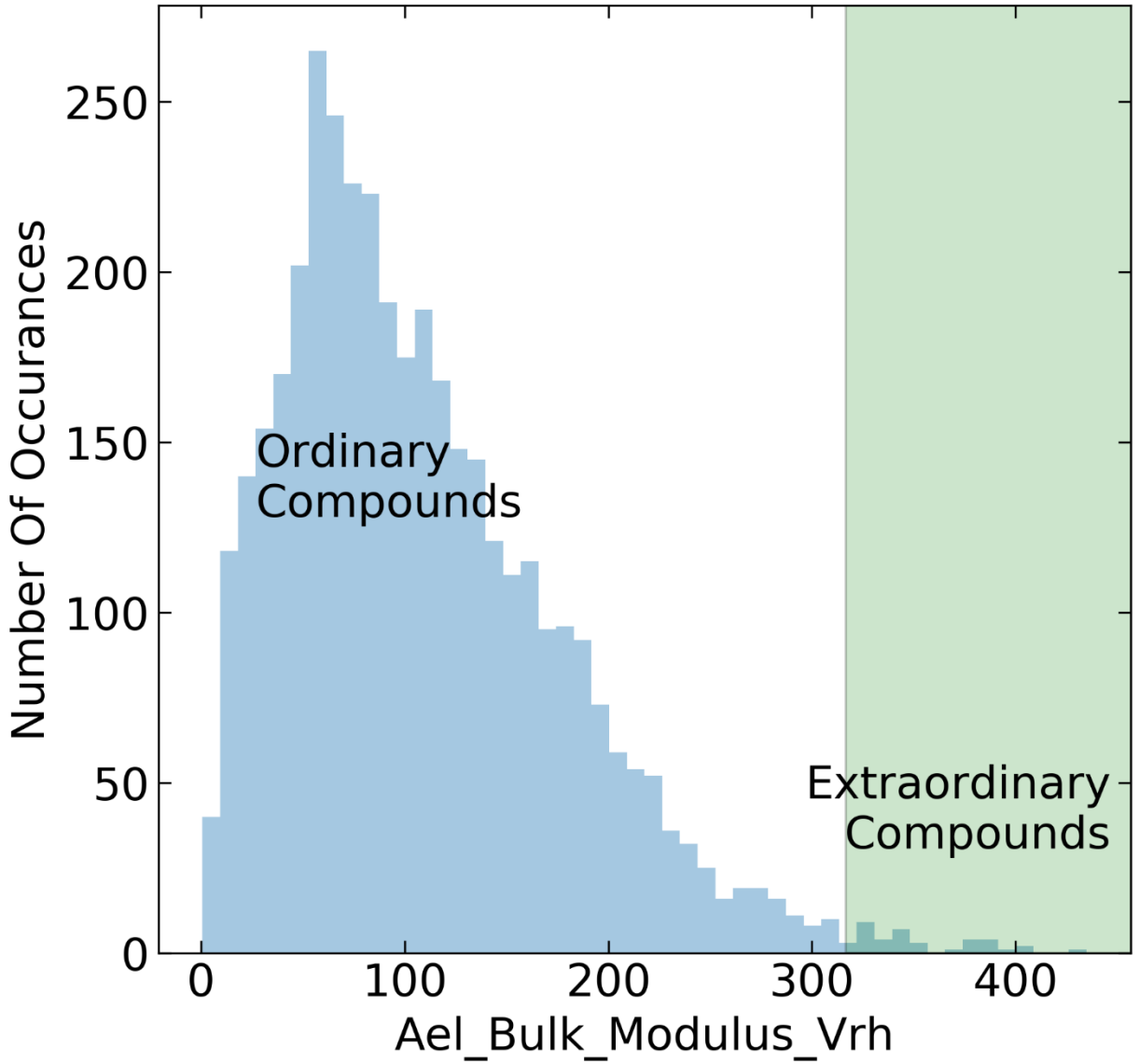
Most machine learning approaches fail with extrapolation





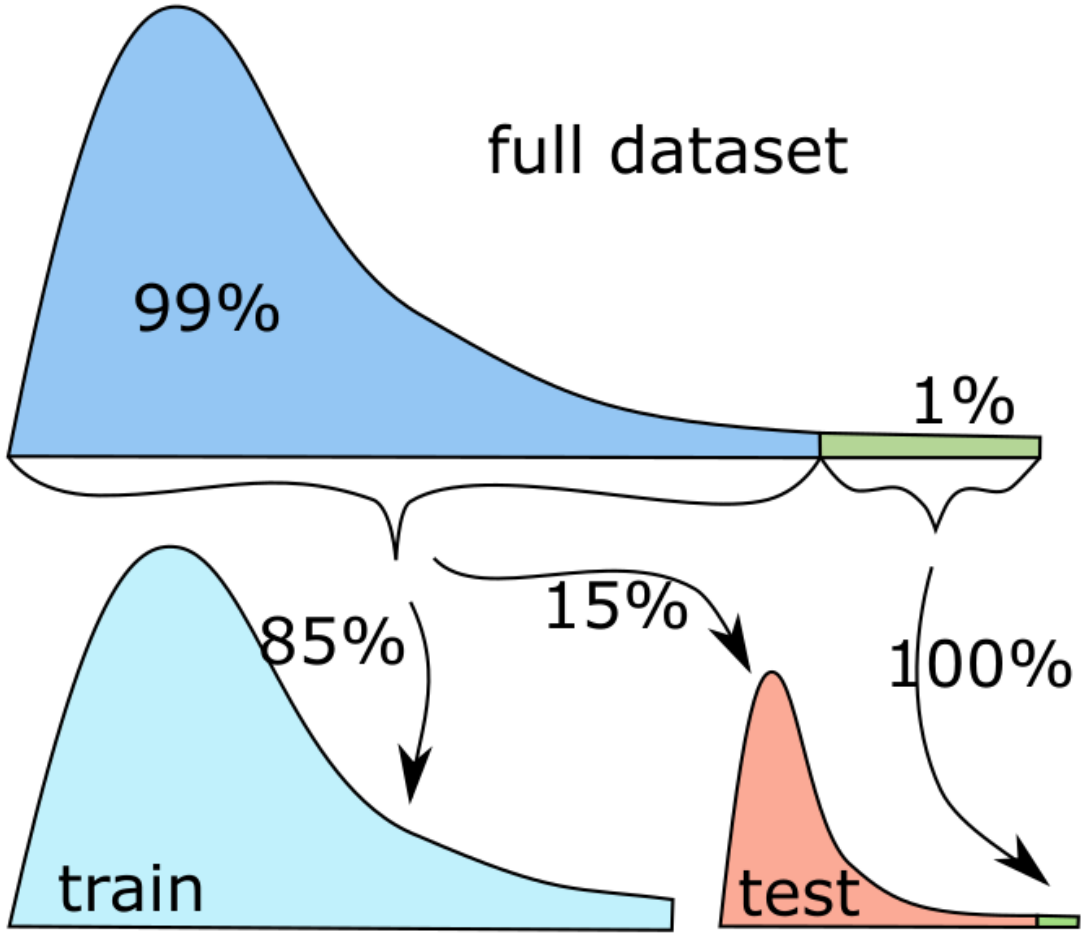
Kauwe, Graser, Murdock, Sparks *Computational Materials Science* (2020)

Will machine learning be able to identify extraordinary compounds?



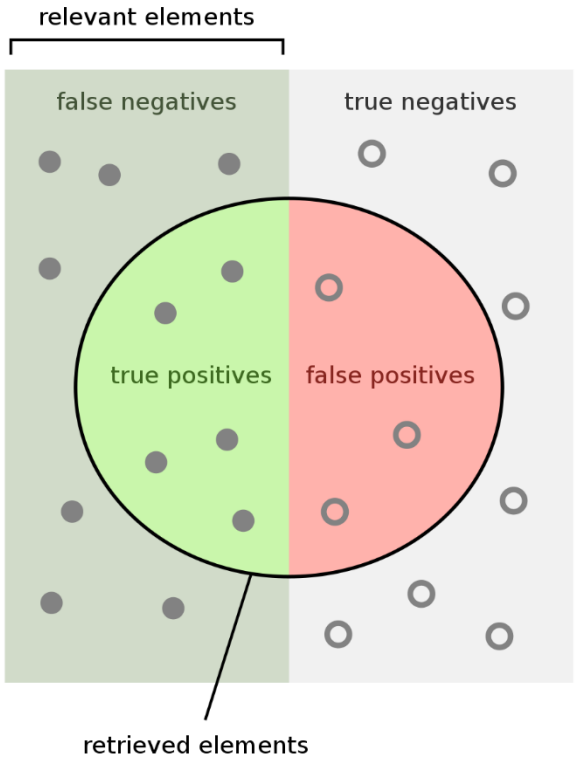
Kauwe, Graser, Murdock, Sparks *Computational Materials Science* (2020)

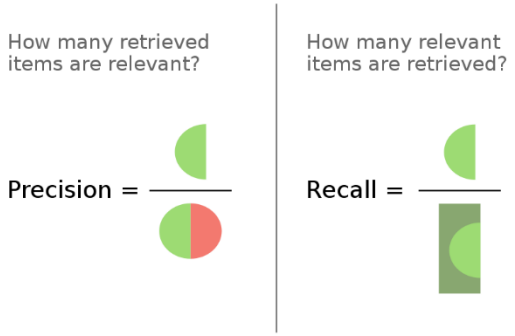
We can test to see how well extrapolation works



Kauwe, Graser, Murdock, Sparks *Computational Materials Science* (2020)

We first need to define precision vs recall





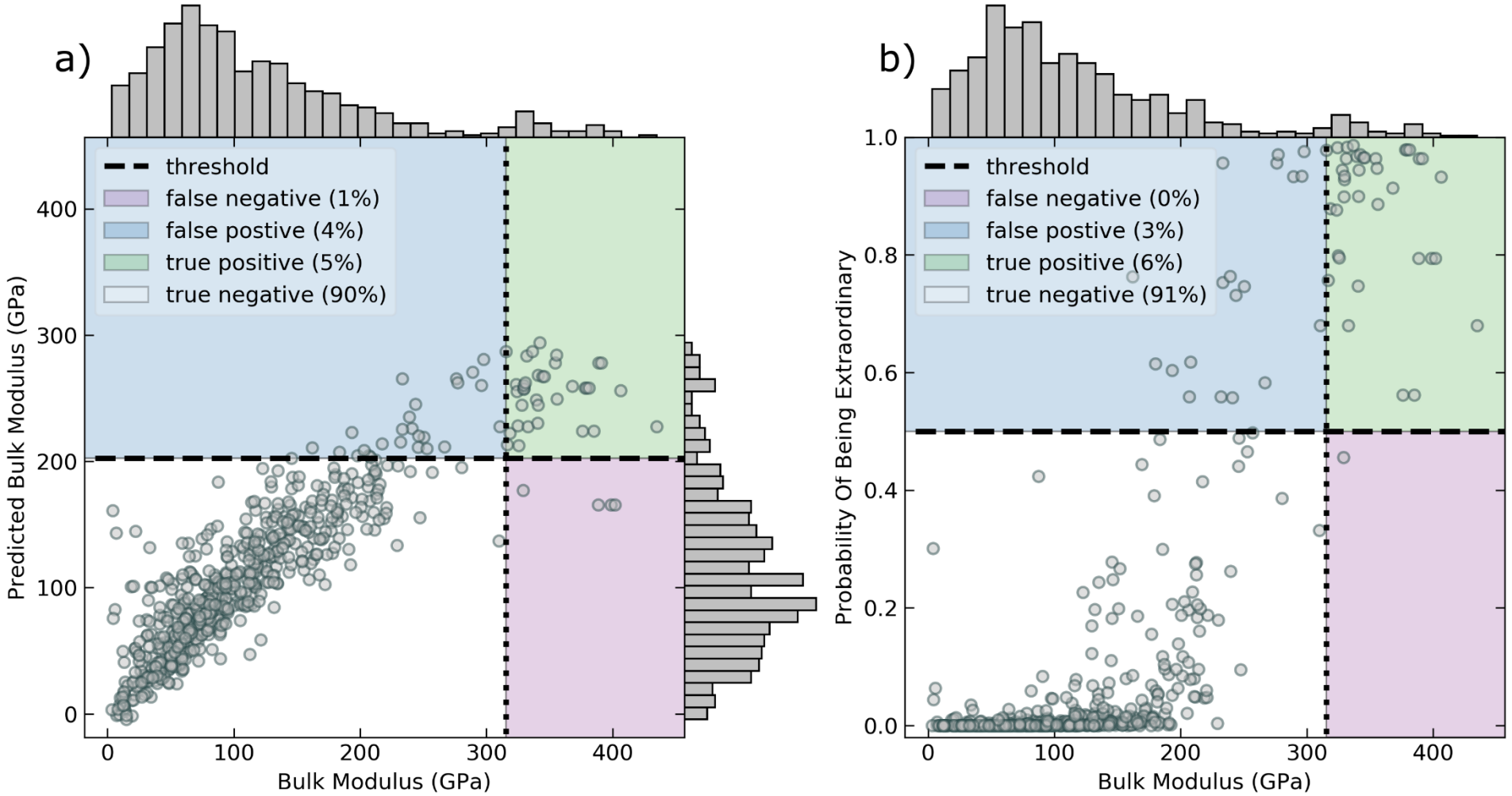
We need to account for both precision and recall!

If we take their harmonic mean, we get F1 score

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

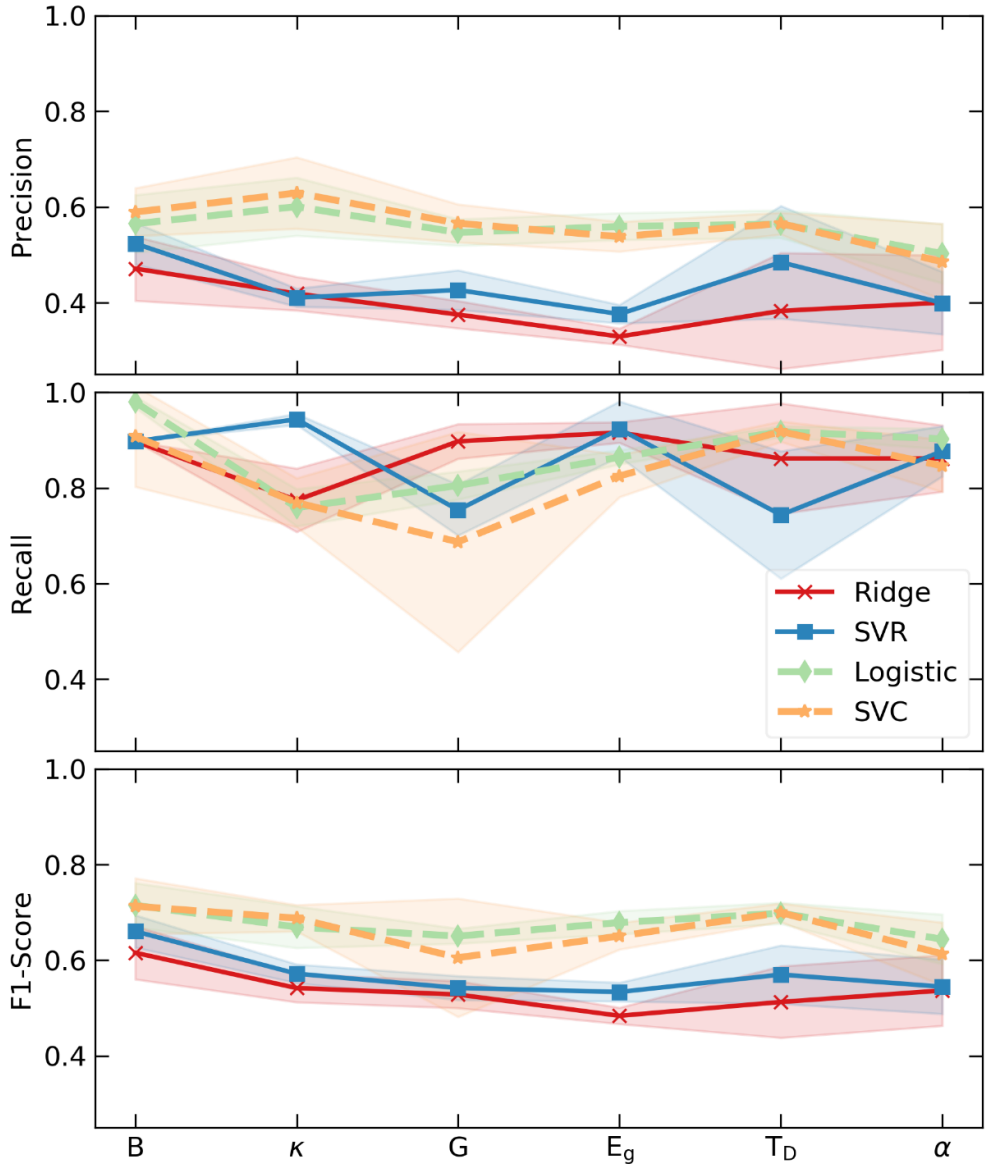
https://en.wikipedia.org/wiki/Precision_and_recall

Precision and recall both suggest extrapolation is possible



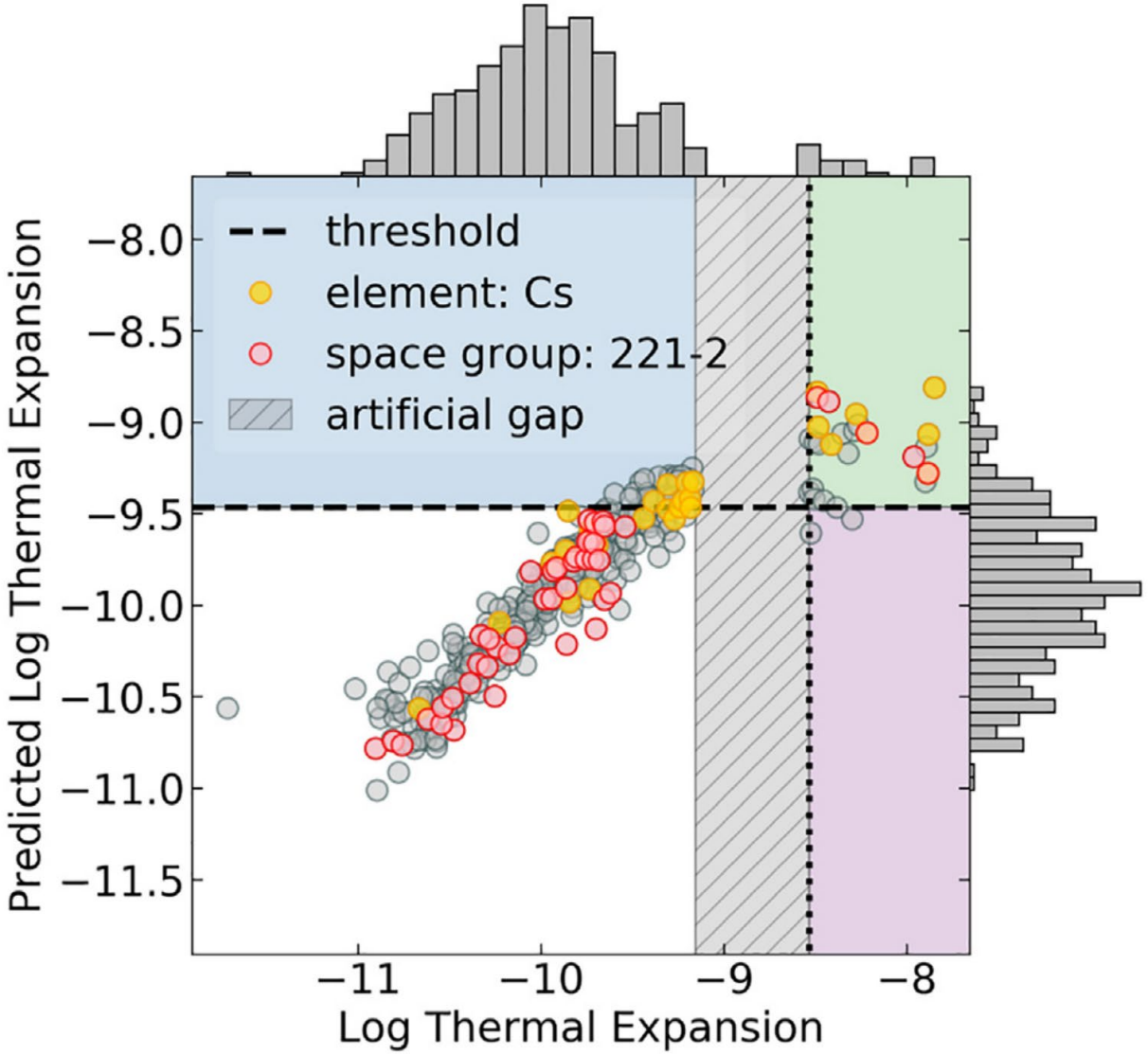
Kauwe, Graser, Murdock, Sparks *Computational Materials Science* (2020)

Classification outperforms regression on almost all properties!



Kauwe, Graser, Murdock, Sparks *Computational Materials Science* (2020)

Extrapolation goes beyond target values!



Kauwe, Graser, Murdock, Sparks *Computational Materials Science* (2020)

Extrapolation goes beyond target values!

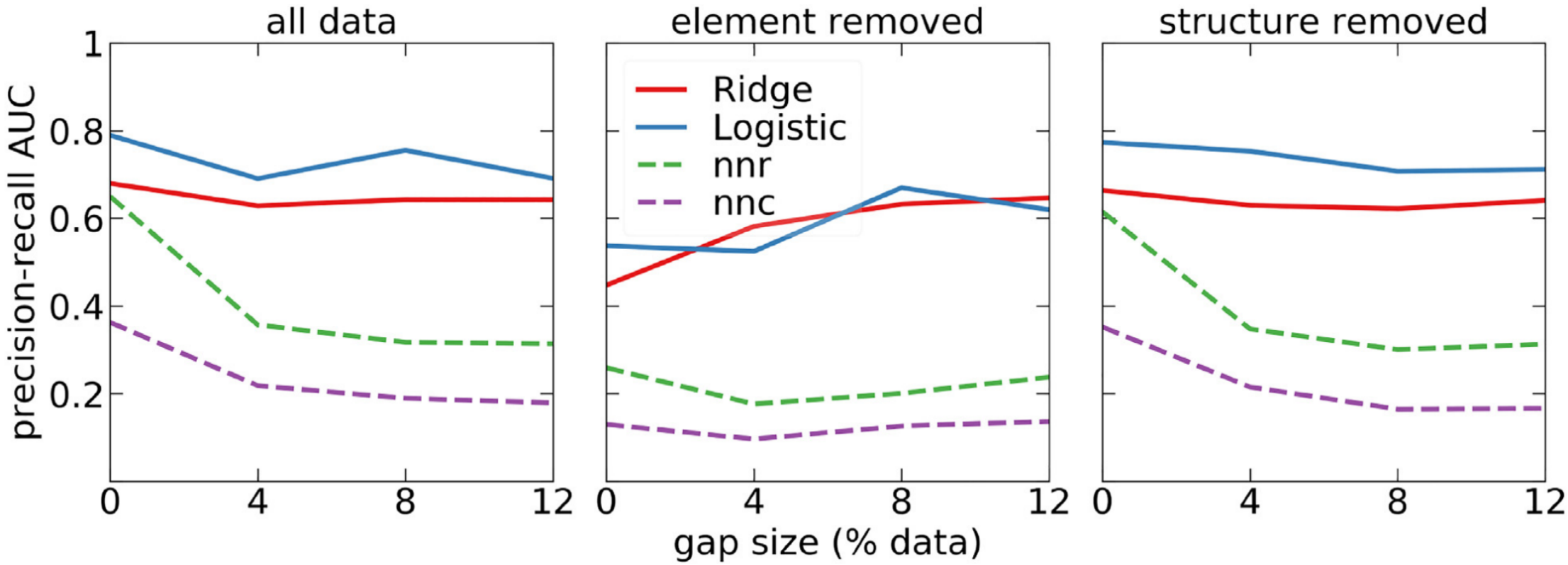


Fig. 5. The precision-recall Area-under-curve (AUC) is computed as the average across all properties. The three extrapolation cases are considered for all gap sizes using four different model types, ridge regression (Ridge), logistic regression (Logistic), nearest-neighbor regression (nnr), and nearest-neighbor classification (nnc). *(left)* The entire training data is used (except the gap), *(center)* the most prevalent element is removed from training, and *(right)* the most prevalent structure is removed from training.

Kauwe, Graser, Murdock, Sparks *Computational Materials Science* (2020)

Compare DFT to experiment to ask if I.I.D. is valid?

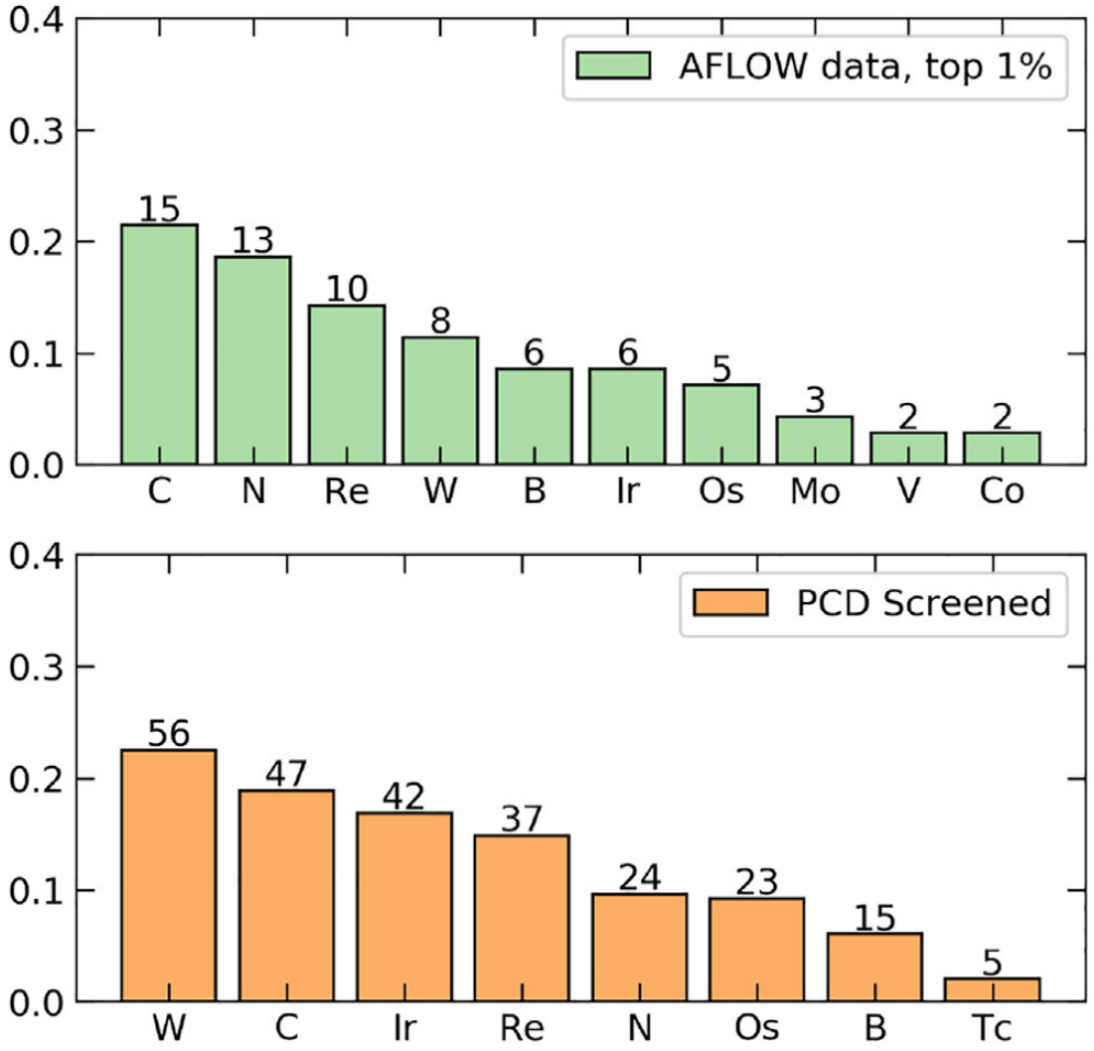


Fig. 6. The ranked elemental prevalence of compounds labeled extraordinary in the original AFLOW data versus those screened from the PCD.

Kauwe, Graser, Murdock, Sparks *Computational Materials Science* (2020)

Clustering