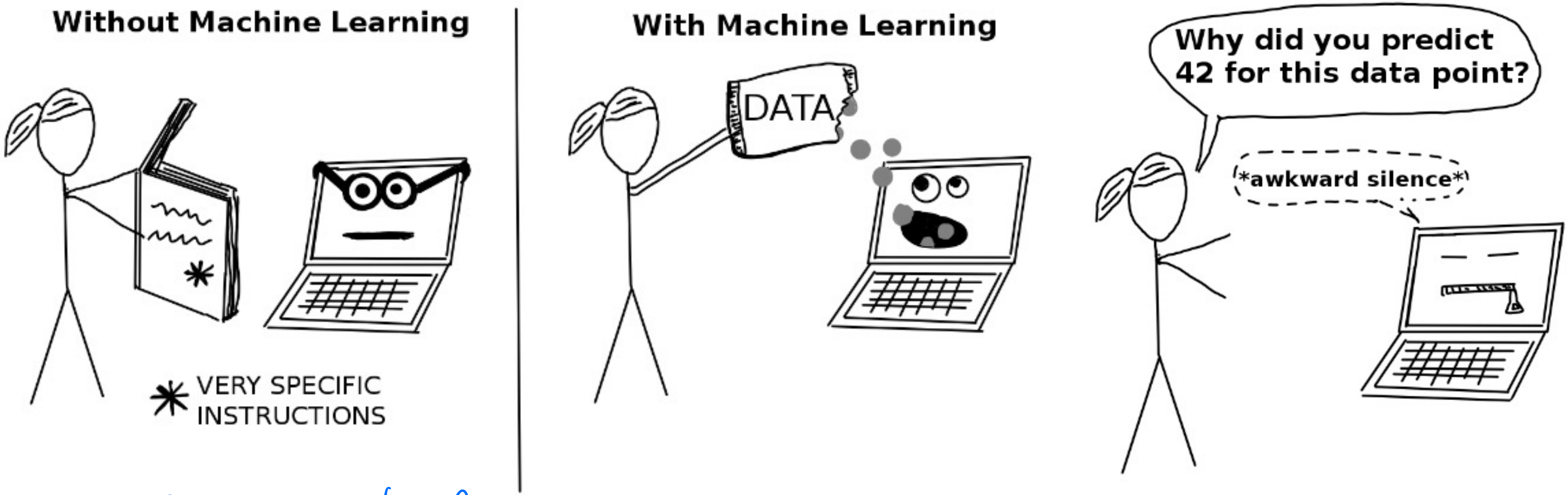
General approaches to interpretable ML

CEMS
Chemical Engineering
& Materials Science



Supervised learning

$$y \approx \hat{y} = f(\vec{x})$$

Why do we care about interpretability?

ACS Partner Journal
ACCOUNTS
of materials research
pubs.acs.org/amrcda
Article

Interpretable and Explainable Machine Learning for Materials Science and Chemistry

Felipe Oviedo,*,# Juan Lavista Ferres, Tonio Buonassisi, and Keith T. Butler*,#

Perspective | Published: 17 March 2022

Interpretable machine learning for knowledge generation in heterogeneous catalysis

Jacques A. Esterhuizen, Bryan R. Goldsmith & Suljo Linic

Nature Catalysis **5**, 175–184 (2022) | Cite this article

7787 Accesses | **21** Citations | **9** Altmetric | Metrics

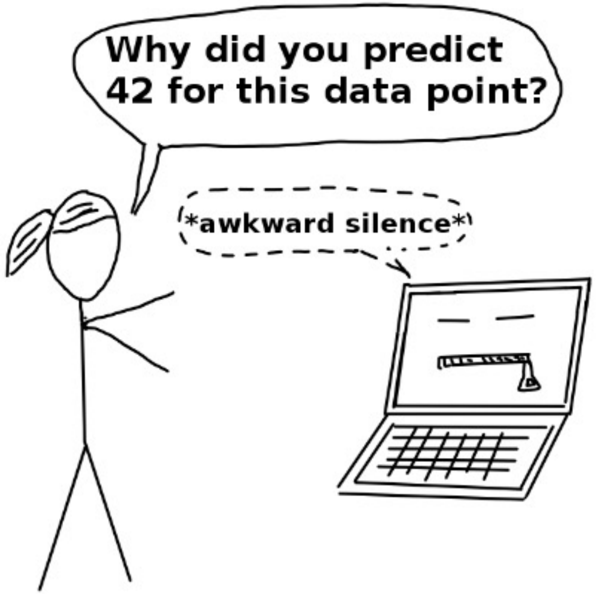
Review Article | Published: 13 October 2020

Drug discovery with explainable artificial intelligence

José Jiménez-Luna, Francesca Grisoni & Gisbert Schneider

Nature Machine Intelligence **2**, 573–584 (2020) | Cite this article

47k Accesses | **207** Citations | **95** Altmetric | Metrics



Why do we care about interpretability?

CEMS Chemical Engineering & Materials Science

ACS Partner Journal

ACCOUNTS —of materials research—

pubs.acs.org/amrcda Article

Interpretable and Explainable Machine Learning for Materials Science and Chemistry

Felipe Oviedo,*,# Juan Lavista Ferres, Tonio Buonassisi, and Keith T. Butler*,#

Perspective | Published: 17 March 2022

Interpretable machine learning for knowledge generation in heterogeneous catalysis

Jacques A. Esterhuizen, Bryan R. Goldsmith & Suljo Linic

Nature Catalysis **5**, 175–184 (2022) | Cite this article

7787 Accesses | **21** Citations | **9** Altmetric | Metrics

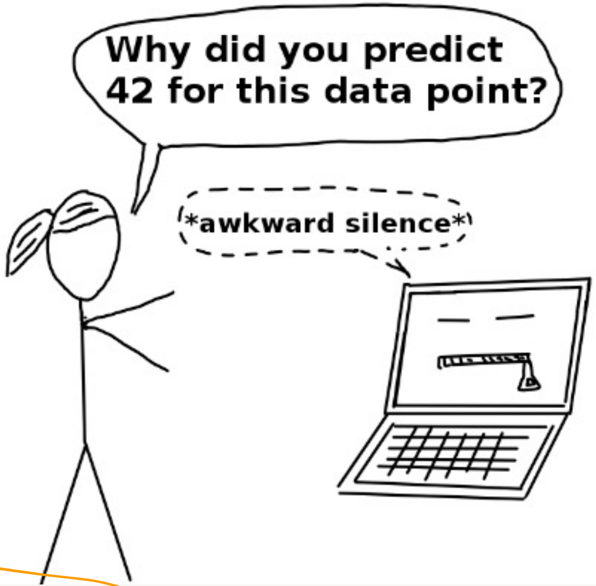
Review Article | Published: 13 October 2020

Drug discovery with explainable artificial intelligence

José Jiménez-Luna, Francesca Grisoni & Gisbert Schneider

Nature Machine Intelligence **2**, 573–584 (2020) | Cite this article

47k Accesses | **207** Citations | **95** Altmetric | Metrics



JOIN CHIME

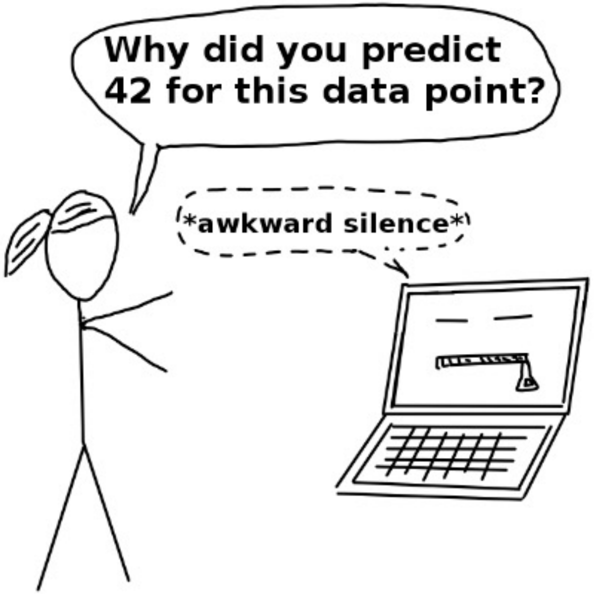
Go to:
https://chimein2.cla.umn.edu/join/608843
Or: visit chimein2.cla.umn.edu and enter
608-843

Why do we care about interpretability?

Reliability → small change in x^j shouldn't lead to large changes in $\hat{y}$

Causality → as x^j changes, can we anticipate $\Delta \hat{y}$

Trust → adoption $\propto$ understanding



Taxonomy of interpretability

Intrinsic → models that are simple
(have few parameters)

Post-hoc → analyze after training
(any black box model)

Local → why was a prediction made ($\hat{y}_i$)

Global → understand general behavior over domain ($\vec{x}$)

Intrinsic interpretability in linear regression

$$\hat{y}(w, x) = w_0 + w_1 x_1 + \ldots + w_p x_p \qquad \min_{w} |Xw - y||_2^2 + \alpha||w||_1$$

Assumptions:
↳ linear relationship b/t y & xj
↳ residuals are normally distributed
↳ features are independent

Interpretation:
↳ $|w_j x_j|$ = feature effect
↳ magnitude of each feature effect is intrinsically important

Intrinsic interpretability in linear regression

CEMS
Chemical Engineering
& Materials Science

$$\hat{y}(w, x) = w_0 + w_1 x_1 + \ldots + w_p x_p \qquad \min_{w} ||Xw - y||_2^2 + \alpha ||w||_1$$

Advantages:

↳ simple

↳ easy to fit

Disadvantages:

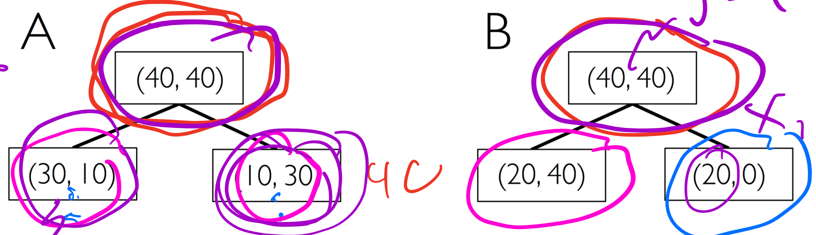
↳ miss nonlinearities

↳ may be inaccurate (at low dimension)

↳ confused by correlated features

Intrinsic interpretability w/ decision trees

CEMS
Chemical Engineering
& Materials Science



$y = (40, 40)$

$x_1 < 3$

40

$\text{Impurity} = -\sum_{i=0}^{N} p(i)\ln(p(i))$ for N classes

$y = [0, 0, \ldots, 1, 0, \ldots]$

$I_0 = -2(0.5\ln 0.5) = 0.69$

$I_A = -\left(\frac{3}{4}\ln\frac{3}{4} + \frac{1}{4}\ln\frac{1}{4}\right) = 0.56$

$I_B = -\frac{3}{4}\left(\frac{1}{3}\ln\frac{1}{3} + \frac{2}{3}\ln\frac{2}{3}\right) = 0.48$

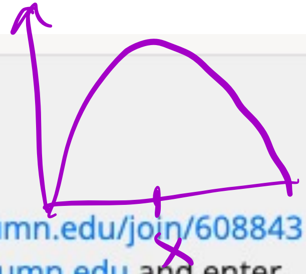
0.13

0.21

39, 40

1, 0

$x\ln x + (1-x)\ln(1-x)$



Information gain
(weighted impurity decrease)

JOIN CHIME

Go to:
https://chimein2.cla.umn.edu/join/608843
Or: visit chimein2.cla.umn.edu and enter
608-843

A dataset to play with

CEMS Chemical Engineering & Materials Science

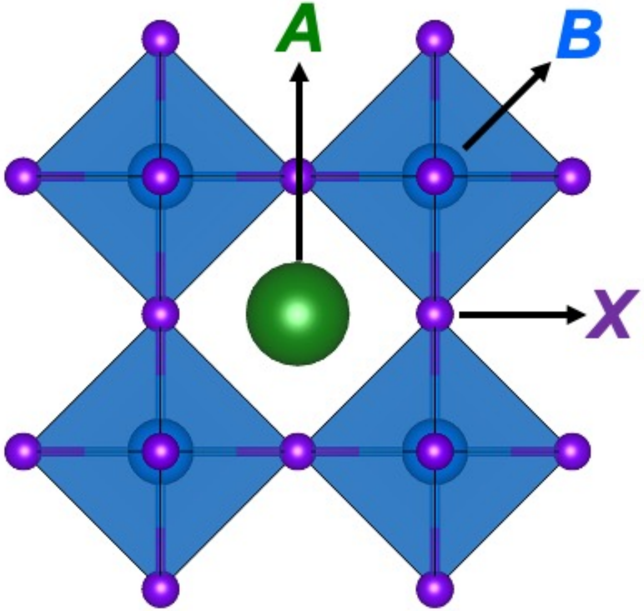
SCIENCE ADVANCES | RESEARCH ARTICLE

MATERIALS SCIENCE

New tolerance factor to predict the stability of perovskite oxides and halides

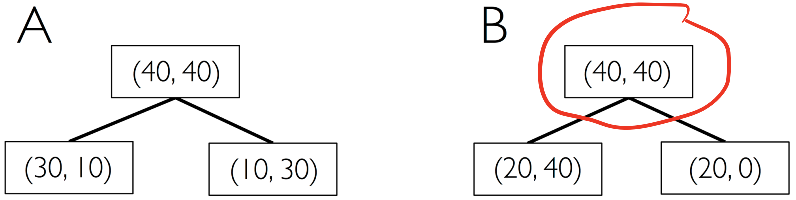
Christopher J. Bartel[1]*, Christopher Sutton[2], Bryan R. Goldsmith[3], Runhai Ouyang[2], Charles B. Musgrave[1,4,5], Luca M. Ghiringhelli[2]*, Matthias Scheffler[2]

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



Intrinsic interpretability w/ tree-based methods

CEMS
Chemical Engineering
& Materials Science



Advantages:

↳ natural (visual) representation

↳ allows for "what-if" explanations.

Disadvantages:

↳ can be too complex

↳ linear relationships

↳ unstable

From model-specific to model-agnostic methods

Model-specific

↳ linear reg ($|w_j x_j|$)

↳ decision trees (IG)

Model-agnostic

↳ interpretations based on predictions (by an arbitrary model)

↳ Global → describe avg behavior

↳ Local → explains individual predictions

One global method: permutation importances

Idea: how much worse our model gets when we randomly shuffle a feature?

1) Train a model, $\hat{y} = f(\vec{x})$
2) Compute the error, $||y - \hat{y}||^2$ (e_{orig})
3) For feature$_j$ in X:
    a) permute x_j
    b) compute, e_j
4) Feature importance, $PFI = \left| \frac{e_j}{e_{orig}} \right|$

$y = 3x_1 + 2x_2^2$

Fisher, Aaron, Cynthia Rudin, and Francesca Dominici. "All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously." http://arxiv.org/abs/1801.01489 (2018)

One global method: permutation importances

CEMS Chemical Engineering & Materials Science

Compute on training data or validation data?

Training → how does training manifest in dependencies between $||y-\hat{y}||^2$ & X^j

Validation → how essential is each feature for generalization?

Fisher, Aaron, Cynthia Rudin, and Francesca Dominici. "All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously." http://arxiv.org/abs/1801.01489 (2018)

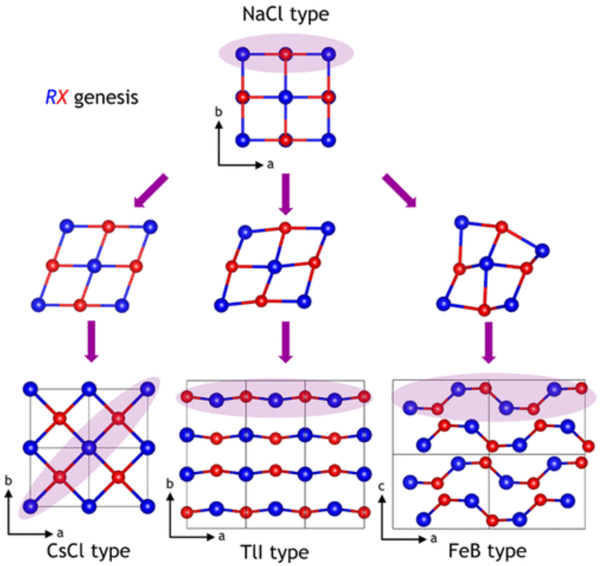
One global method: permutation importances

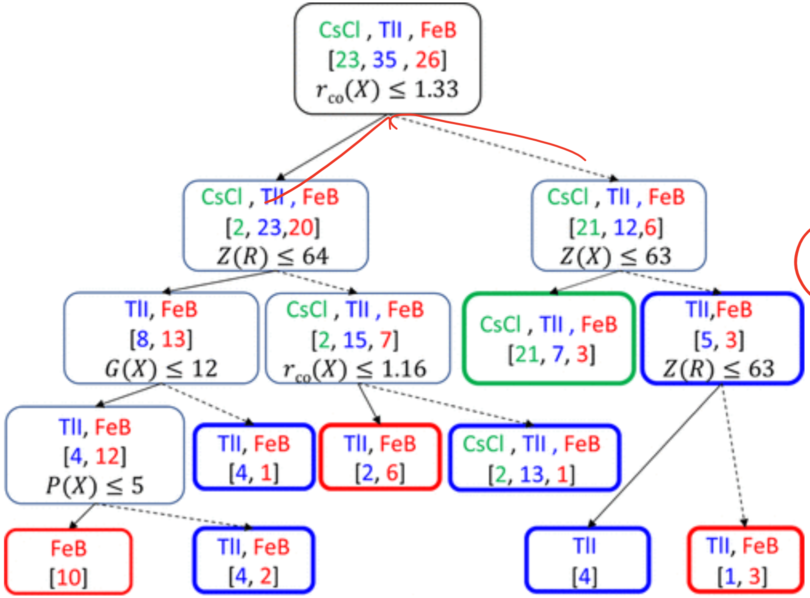
CHEMISTRY OF MATERIALS
pubs.acs.org/cm
Article

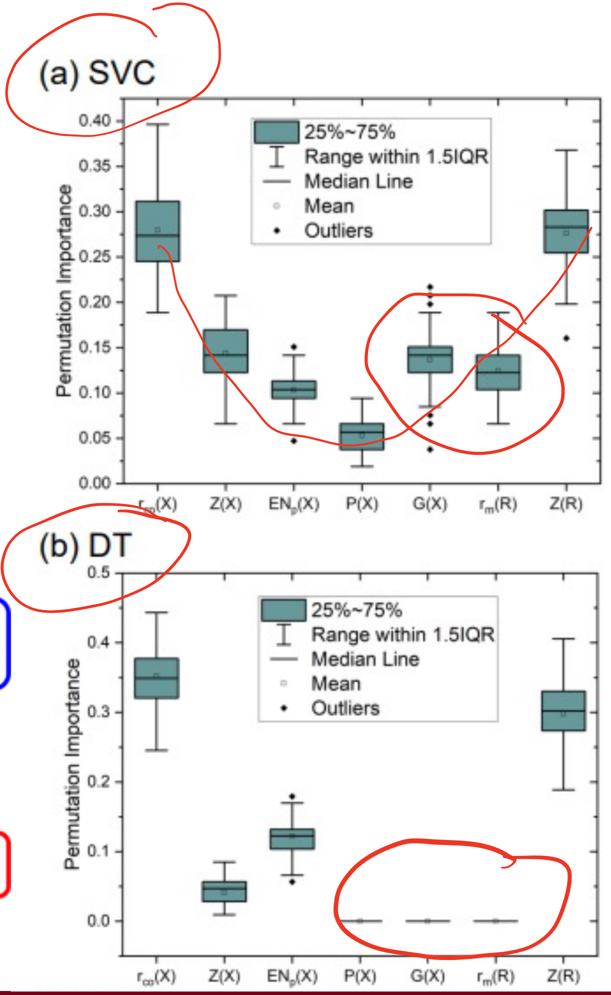
Revealing Hidden Patterns through Chemical Intuition and Interpretable Machine Learning: A Case Study of Binary Rare-Earth Intermetallics *RX*

Volodymyr Gvozdetskyi,* Balaranjan Selvaratnam, Anton O. Oliynyk, and Arthur Mar*

Cite This: *Chem. Mater.* 2023, 35, 879−890
Read Online







One global method: permutation importances

Advantages:

↳ easy to interpret

↳ does not require re-training

Disadvantages:

↳ permutations may lead to unphysical combinations of features

↳ to get robust estimates, we need many repeats

Fisher, Aaron, Cynthia Rudin, and Francesca Dominici. "All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously." http://arxiv.org/abs/1801.01489 (2018)

Quick survey of other global methods

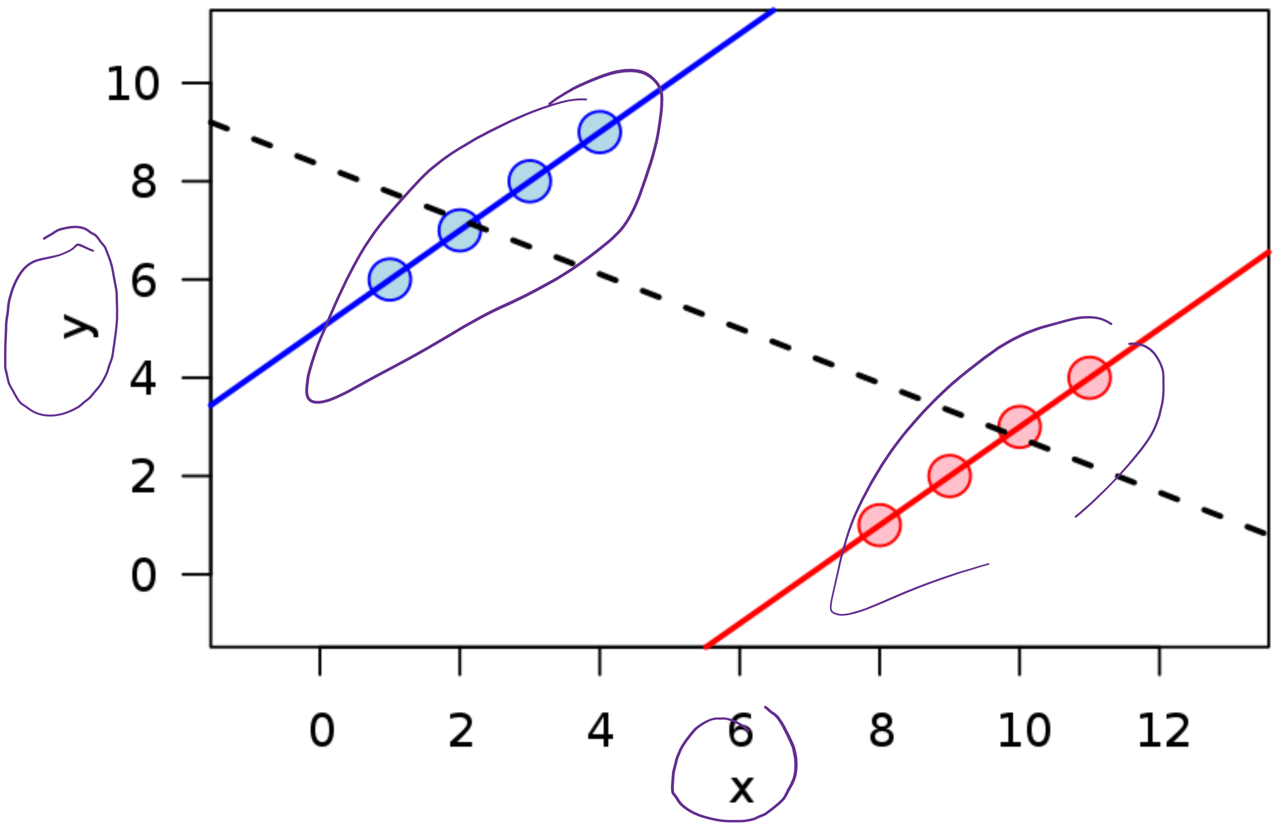
CEMS
Chemical Engineering
& Materials Science

Partial dependencies
↳ how does $\hat{y}$ vary wrt a single feature, x^j

Feature interaction
↳ how does the value of one feature influence the partial dependence of another feature.

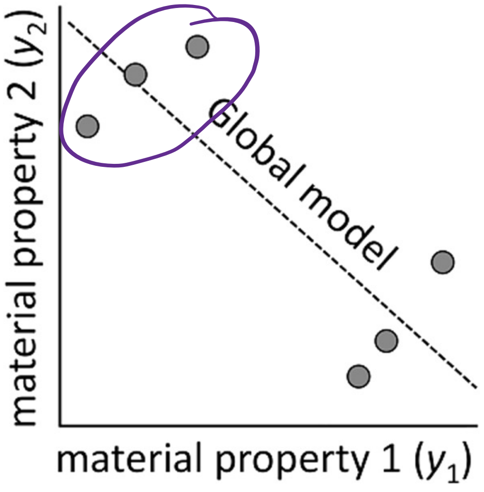
Surrogate model
↳ train an intrinsically model on my black box model's predictions, then interpret that model

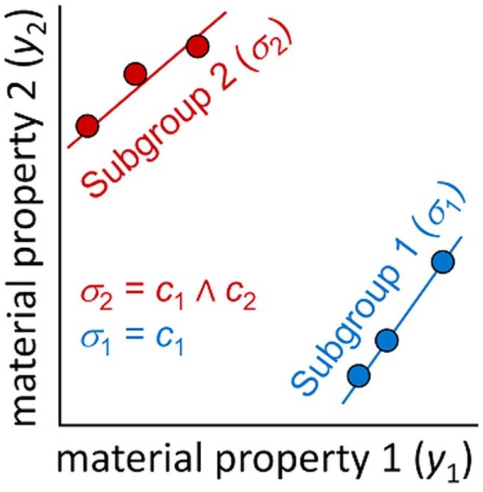
From global to local methods

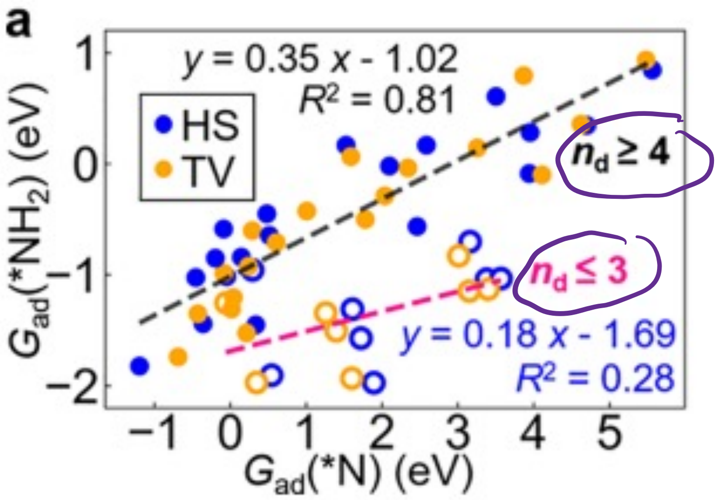


https://en.wikipedia.org/wiki/Simpson's_paradox

Our data may not be globally interpretable





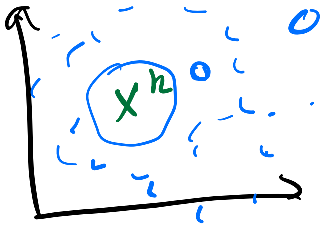


Subgroup discovery → a systematic way to find related "sub" populations

 Goldsmith et al., New Journal of Physics, 2017, 19, 013031 | Li et al., ACS Catalysis, 2021, 11, 7906

Local version of surrogate: LIME

LIME → Local Interpretable Model-agnostic Explanations



Given blackbox model, $\hat{y} = f(\vec{x})$

1) Select a point to probe, x^k
2) Randomly generate new data points
3) Predict target ($\hat{y}$) for new points using $f(\vec{x})$
4) Weight each new pt based on proximity to x^k
5) Train an interpretable model on new pts w/ weighted loss function
6) Interpret

$\alpha_q \|y_i - \hat{y}_i\|^2$

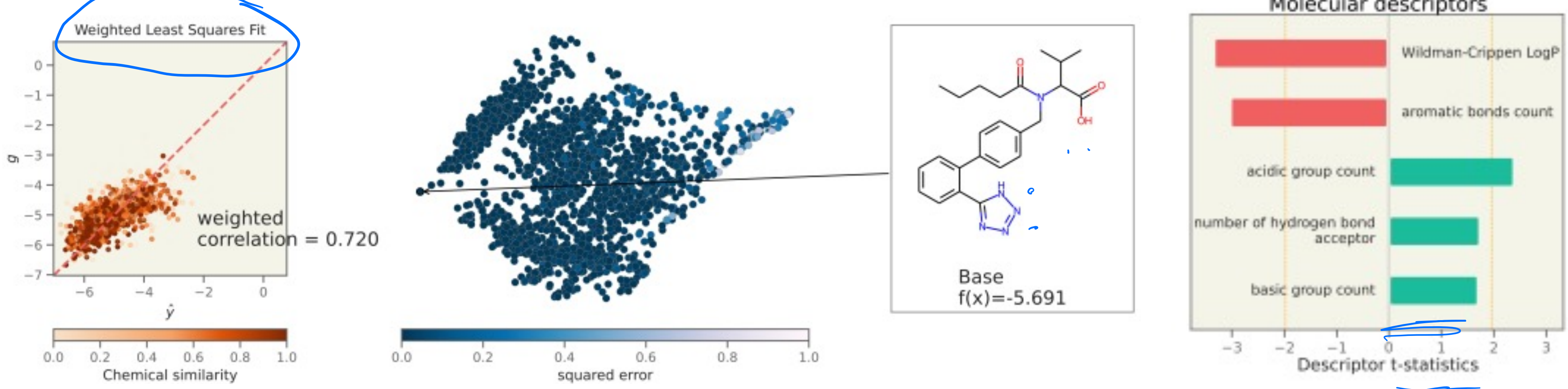
Local version of surrogate: LIME

EXPLAINING MOLECULAR PROPERTIES WITH NATURAL LANGUAGE

A PREPRINT

Heta A. Gandhi
Department of Chemical Engineering
University of Rochester
Rochester, NY, 14627
hgandhi@ur.rochester.edu

Andrew D. White*
Department of Chemical Engineering
University of Rochester
Rochester, NY, 14627
andrew.white@rochester.edu



https://github.com/ur-whitelab/exmol

exmol Public

Another local method: counterfactuals

Check for updates

Cite this: *Chem. Sci.*, 2022, **13**, 3697

All publication charges for this article

Model agnostic generation of counterfactual explanations for molecules†

Geemi P. Wellawatte,[a] Aditi Seshadri[b] and Andrew D. White *[b]

$AsNbO_3$, $X = [r_A, r_B, r_X]$
Δ Δ Δ

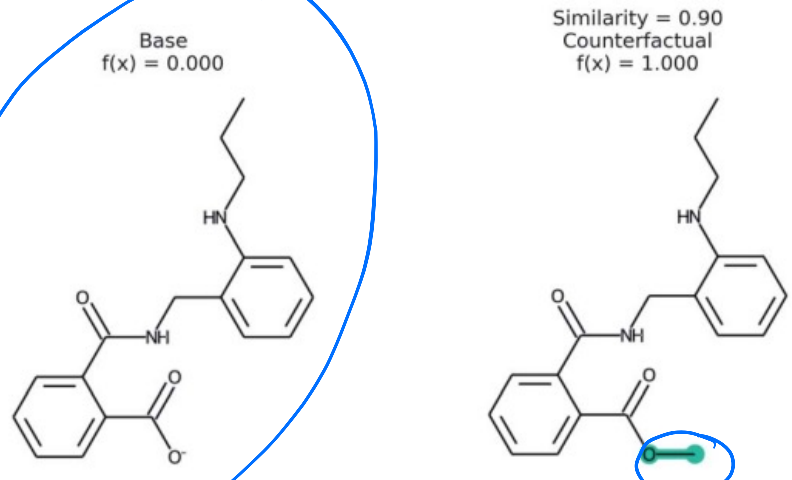


Fig. 1 An example of a counterfactual. The molecule on left was predicted to have class of 0, no activity. With the modification shown in teal, the molecule would be in class 1, active. This shows that the carboxylic acid is an explanation for lack of activity.

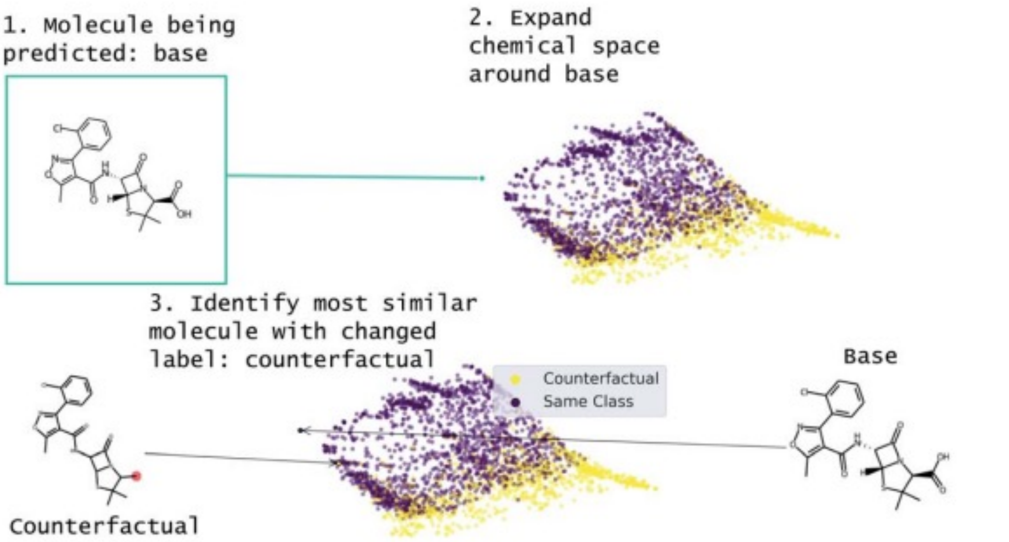


Fig. 2 Overview of MMACE. The input is a molecule to be predicted. Chemical space is expanded and clustered. Counterfactuals are selected from clusters to find succinct explanation of base molecule prediction.

Idea: Let features play a "game" to determine feature effects

Given: $\hat{y} = f(\vec{x})$

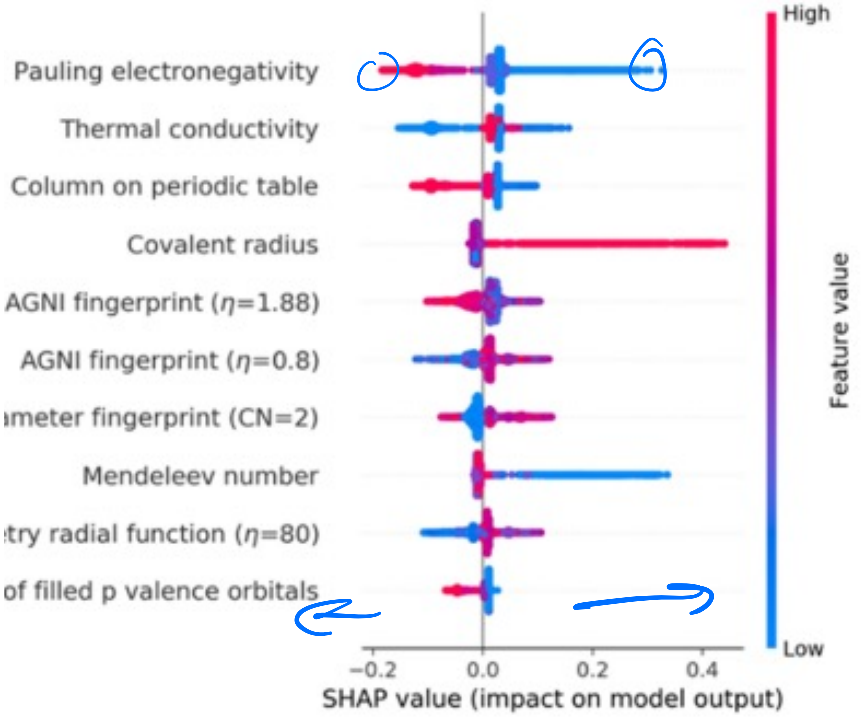
Question: How much does x^j contribute to each pts deviation from the mean

$$\Delta y_i = \langle \hat{y} \rangle - y_i$$

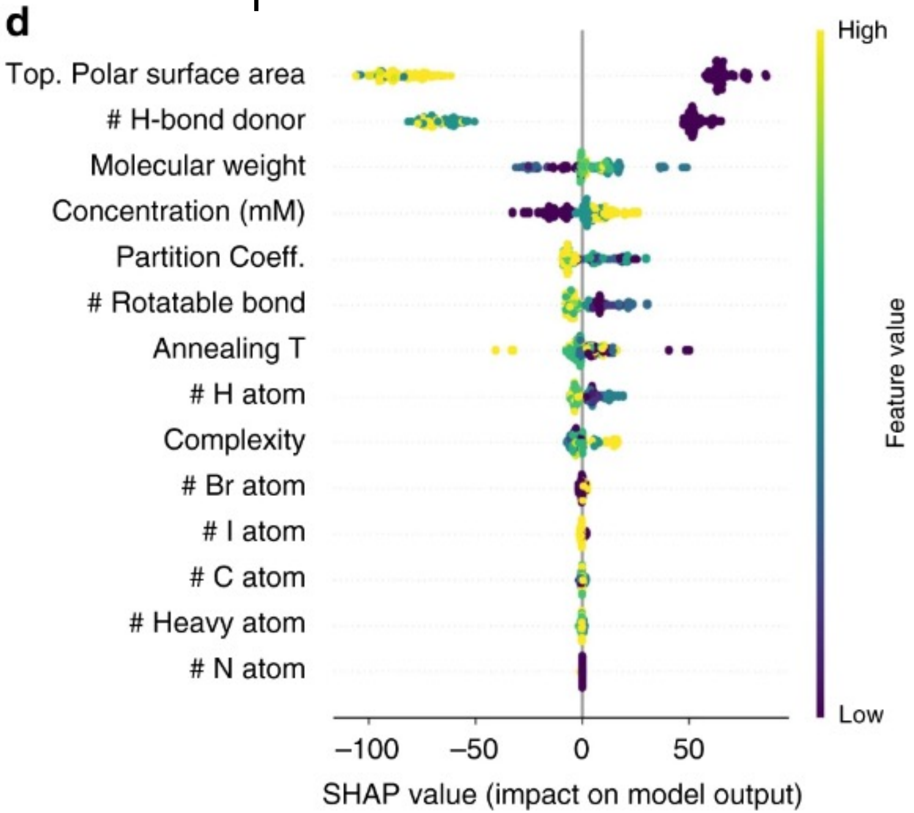
1) Pick a feature & data pt (x^j, y_k)
2) Draw a random subset including x^j (of features)
3) Compute $\hat{y} = f(\tilde{x})$
4) Repeat this but w/o x^j → ≈ marginal contribution to $\hat{y}$
5) Repeat over all features → subsets → data pts

Another local/global method: Shapley values and SHAP

Partial charges in nanoporous materials



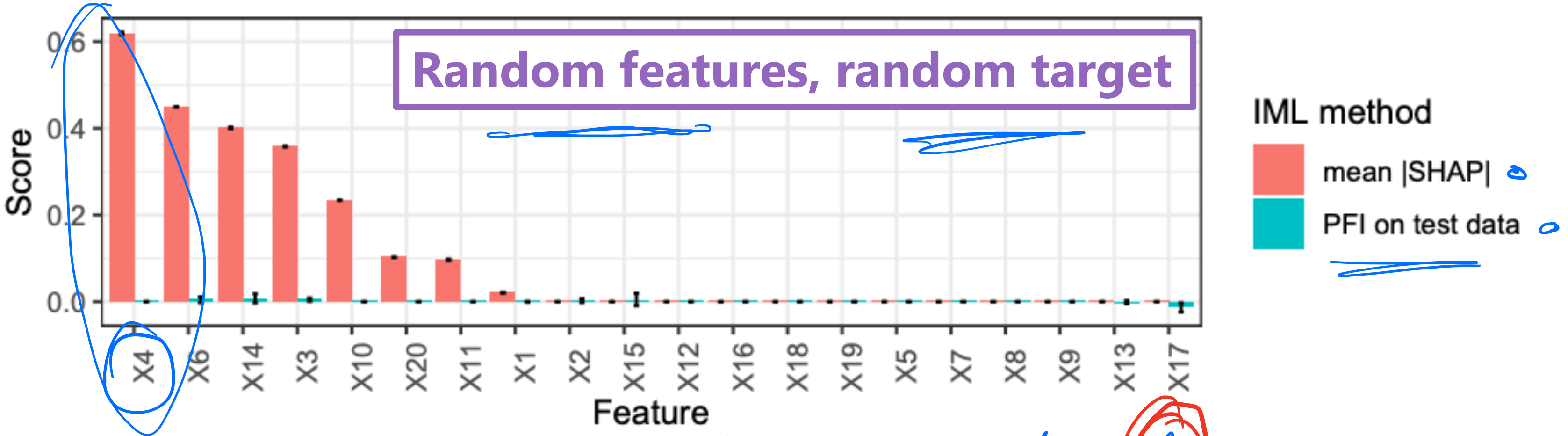
Capping layers for hybrid halide perovskite solar cells



Korlev et al., Chem Mater, 2020, 32, 7822 | Hartono et al, Nature Comm, 2020, 11, 4172

Pitfalls of interpretable ML

1. Assuming one method will always work



SHAP → which x_i contribute $\hat{y}$

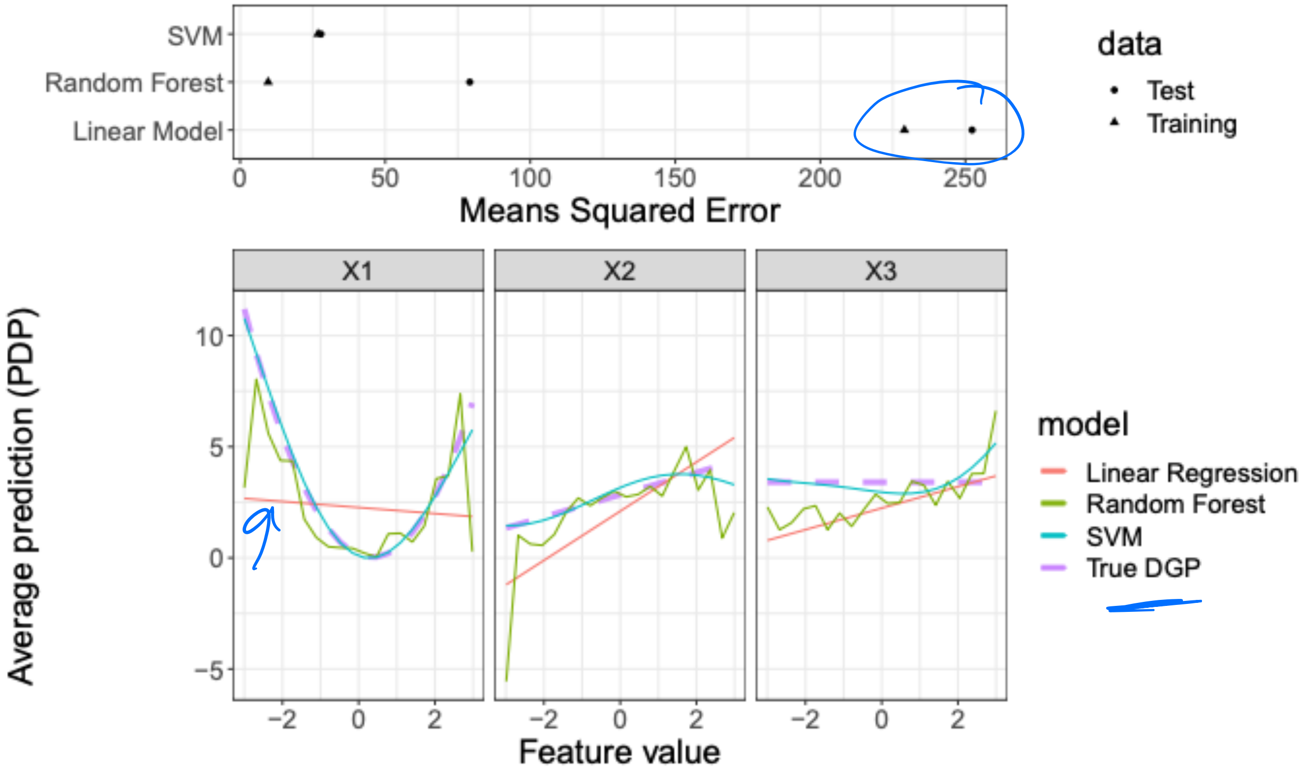
PFI → which x_i control $\|y - \hat{y}\|^2$ (on test)

Molnar et al., 2020, arXiv:2007.04131

Pitfalls of interpretable ML

2. Bad model generalization

GIGO



Molnar et al., 2020, arXiv:2007.04131

Pitfalls of interpretable ML

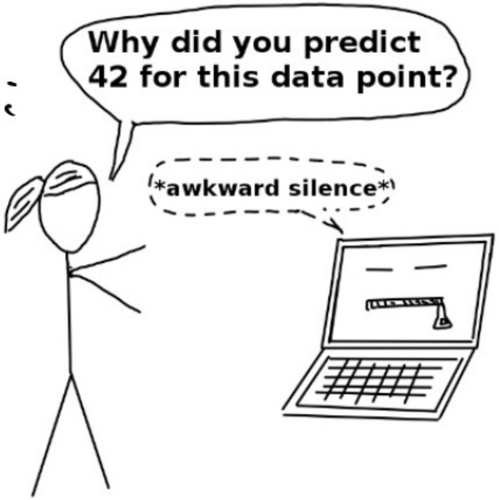
3. Unnecessary complexity

Recall why we care about interpretability:

Reliability → small change in x_i shouldn't lead to a large change in $\hat{y}_i$

Causality → as we change x_i, can we anticipate change in $\hat{y}_i$

Trust ⇒ adoption ∝ understanding



Intrinsic interpretability is always preferred!

Molnar et al., 2020, arXiv:2007.04131