OPTIMIZING QSA/PR MODEL INTERPRETABILITY, STABILITY & PERFORMANCE

NATHANIEL CHAREST, PH.D.

ORD-CCTE-CCED-CCCB

THIS PRESENTATION DOES NOT NECESSARILY REFLECT U.S. EPA POLICY

OVERVIEW

What is a QSA/PR & Why Do We Want to Model It

How Do We Model?

Judging Models: Bias & Variance

Chemical Space, Embeddings & Descriptors

A Posteriori Incantatum: Inferring a Mechanism

External Testing

Reporting

Conclusions

QUANTITATIVE STRUCTURE-ACTIVITY/PROPERTY RELATIONSHIPS

- A mapping between the chemical nature of a substance and its biological or physical action
  - Structure-Activity is a biological action, such presence in blood metabolic half life or the aquatic concentration at which 50% of *P. promelas* dies
  - Structure-Property is a physical quality, such as water solubility or acidity

QUANTITATIVE STRUCTURE-ACTIVITY/PROPERTY RELATIONSHIPS

- A mapping between the chemical nature of a substance and its biological or physical action
  - Structure-Activity is a biological action, such presence in blood metabolic half life or the aquatic concentration at which 50% of *P. promelas* dies
  - Structure-Property is a physical quality, such as water solubility or acidity
- Activities tend to be complex
  - Biological behaviors can manifest through many mechanisms and be derived from many structural sources
  - For instance, a faster metabolic half-life might be due to pi systems attracting the action of liver metabolism, the presence of moieties allowing for fast entry into the Krebs cycle, or both

QUANTITATIVE STRUCTURE-ACTIVITY/PROPERTY RELATIONSHIPS

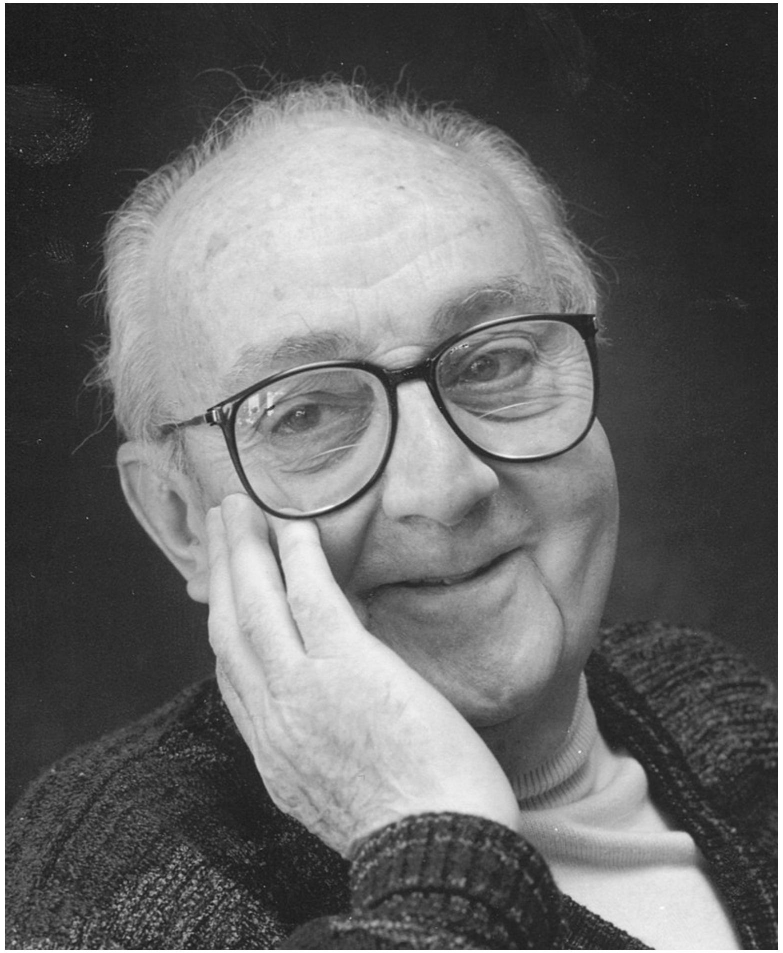
- A mapping between the chemical nature of a substance and its biological or physical action
  - Structure-Activity is a biological action, such presence in blood metabolic half life or the aquatic concentration at which 50% of *P. promelas* dies
  - Structure-Property is a physical quality, such as water solubility or vapor pressure
- Activities tend to be complex
  - Biological behaviors can manifest through many mechanisms and be derived from many structural sources
  - For instance, a faster metabolic half-life might be due to pi systems attracting the action of liver metabolism, the presence of moieties allowing for fast entry into the Krebs cycle, or both
- Properties tend to be simple
  - Water solubility is a function of relatively few molecular interactions

QUANTITATIVE STRUCTURE-ACTIVITY/PROPERTY RELATIONSHIPS

- A mapping between the chemical nature of a substance and its biological or physical action
  - Structure-Activity is a biological action, such presence in blood metabolic half life or the aquatic concentration at which 50% of *P. promelas* dies
  - Structure-Property is a physical quality, such as water solubility or vapor pressure
- Activities tend to be complex
  - Biological behaviors can manifest through many mechanisms and be derived from many structural sources
  - For instance, a faster metabolic half-life might be due to pi systems attracting the action of liver metabolism, the presence of moieties allowing for fast entry into the Krebs cycle, or both
- Properties tend to be simple
  - Water solubility is a function of relatively few molecular interactions
- I'm just going to call it "Structure-Activity" from this point on

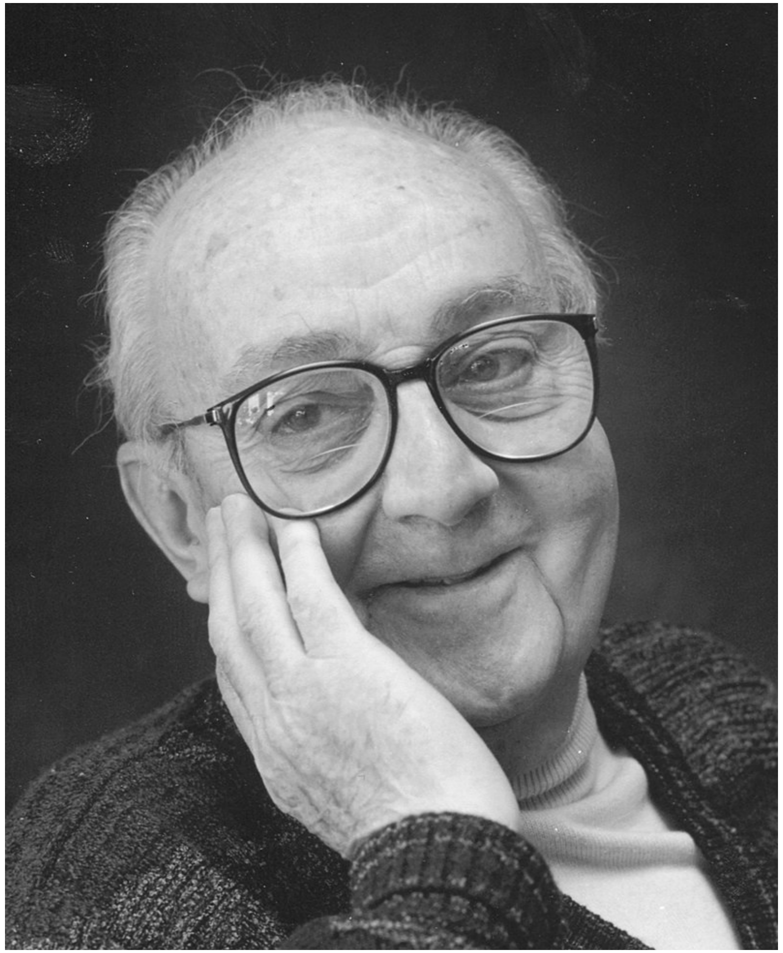
MAIS POURQUOIS?

- Can a model beat a good experiment?
  - "All models are wrong but some are useful." George Box. 1987. *Empirical Model-Building and Response Surfaces.*



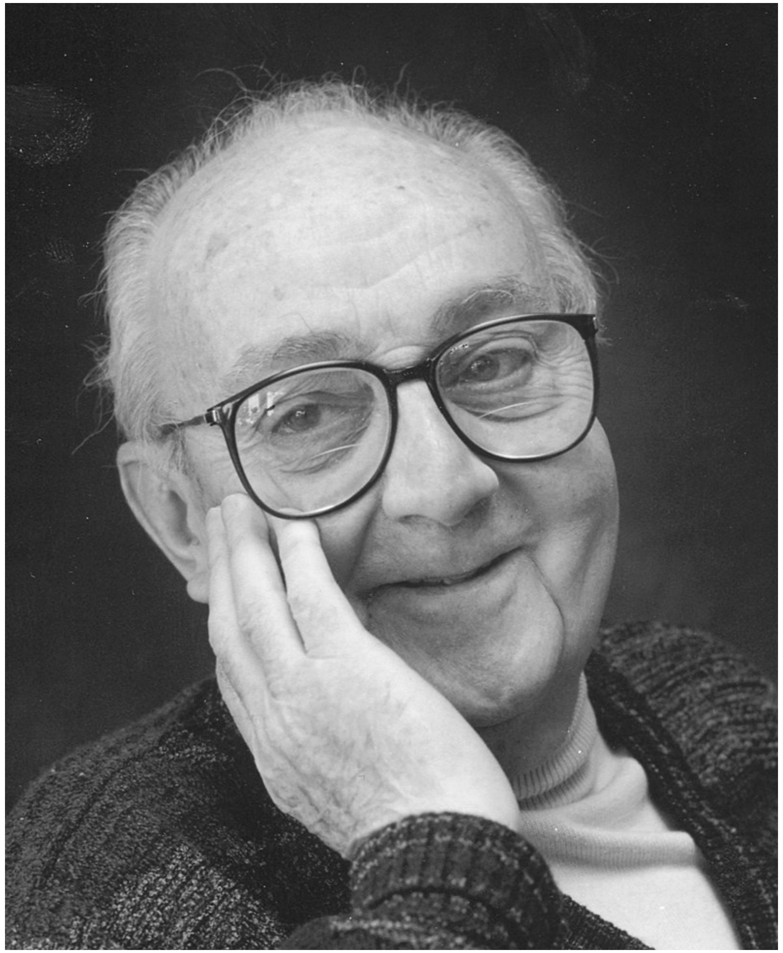
MAIS POURQUOIS?

- Can a model beat a good experiment?
  - "All models are wrong but some are useful." George Box. 1987. *Empirical Model-Building and Response Surfaces.*
- So Why Model?



MAIS POURQUOIS?

- So Why Model?
  - There are a lot of chemicals and experimenting explicitly can be resource prohibitive
    - Models are cheap and reduce *in vivo* testing
  - Experiments are imperfect
    - Models shine light on data points that are mathematically abnormal (random errors) and can detect signals that indicate systematic errors
  - Large Structure-Activity data is structured but often beyond human comprehension
    - Models are mathematical tools to investigate the patterns within data



HOW DO WE MODEL QSA RELATIONSHIPS?

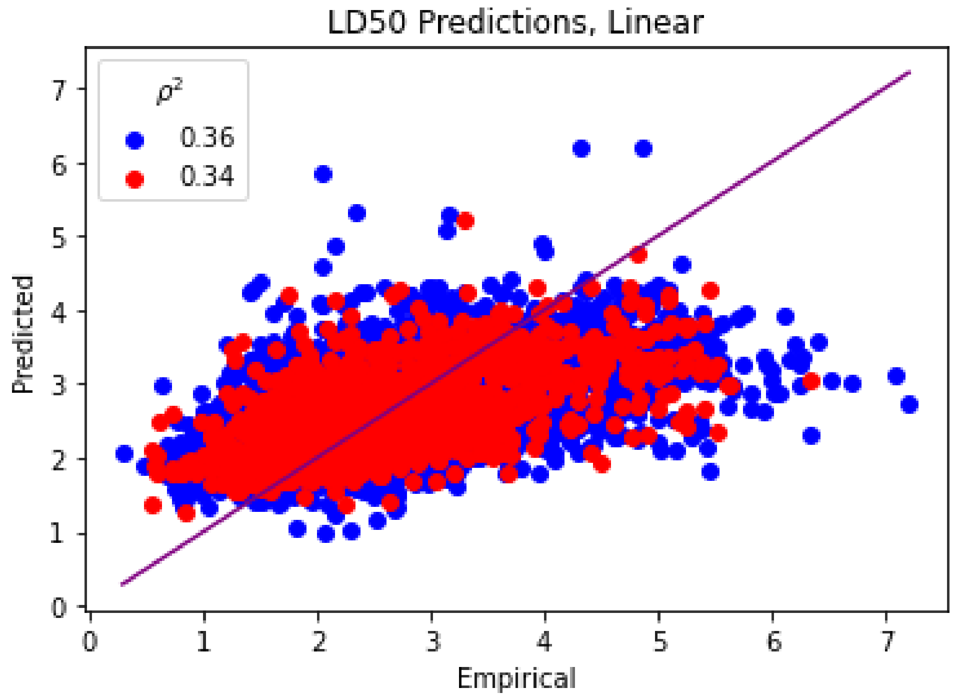
Datasets

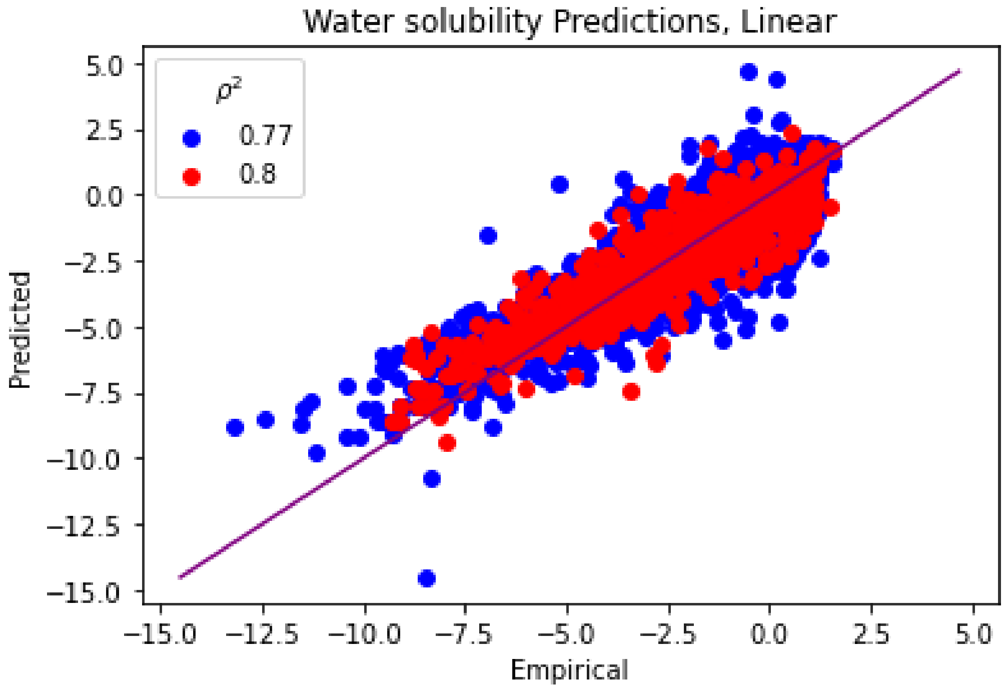
“LD50” – The dosage at which a compound was found to kill 50% of rat model organisms

“Water solubility” – Measure of the amount of a substance that can dissolve in water at a specific temperature

THE OLD WAYS

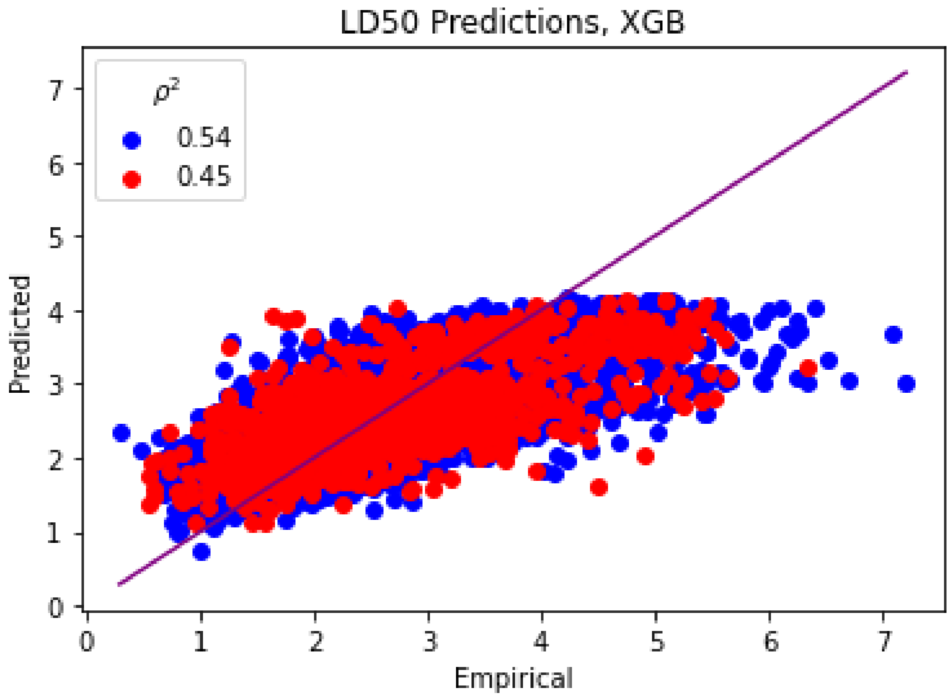
- Historically, QSAR has applied one-line equations with relatively few fitting parameters
  - Linear models
  - Additive Expressions
  - Logistic regression/classification
  - “High bias, low variance”
- These models are simple to calculate
- These models tend to be easily interpretable

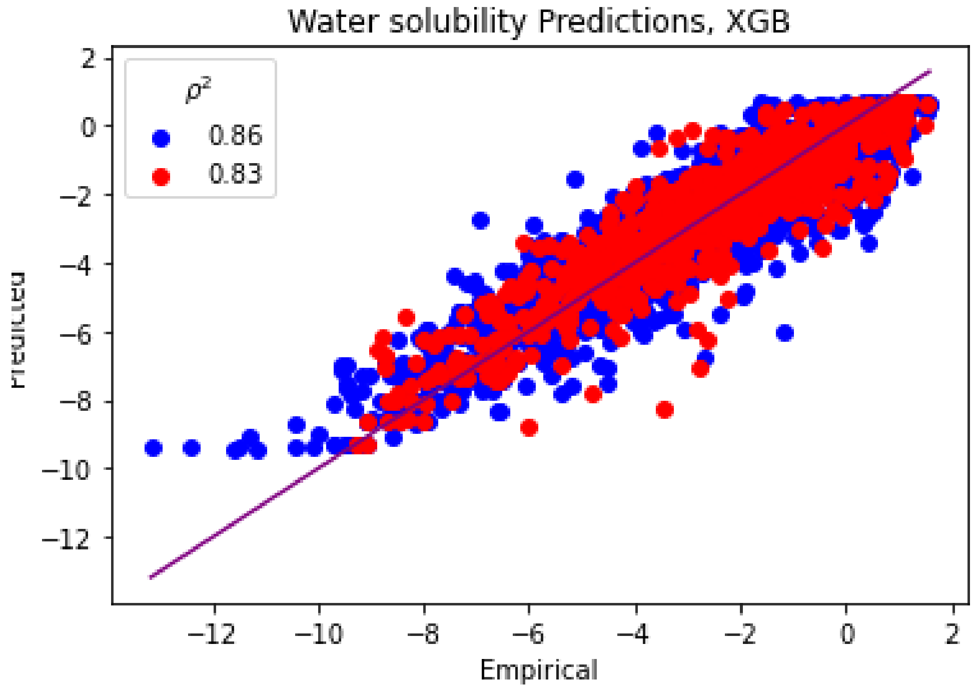

LD50 Predictions, Linear


Water solubility Predictions, Linear

AND THE NEW

- Modern approaches use ‘machine learning methods’, which require larger memory resources, many more fitting parameters, or both
  - K-Nearest Neighbor
  - Decision Trees, Random Forests & Gradient Boosted Trees
  - Support Vector Machines
  - Neural Networks & Representation Learning
  - “Low bias, high variance”
- Computationally expensive
- “Black Box” difficult to interpret





JUDGING A MODEL

MACHINE LEARNING HAS INTRODUCED NEW CONSIDERATIONS

Bias

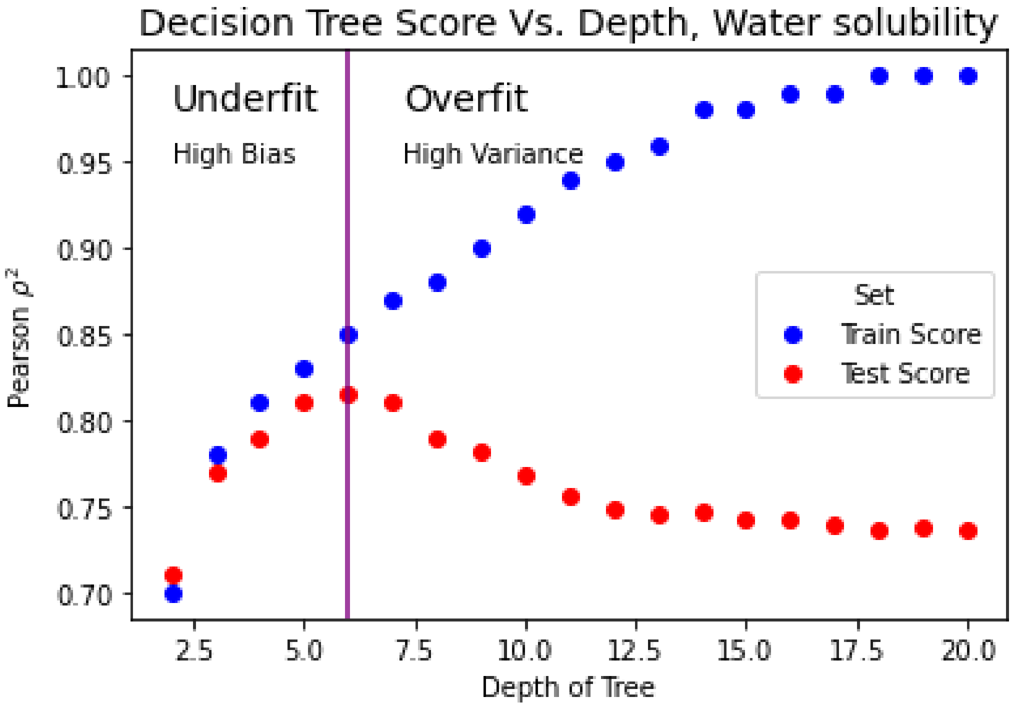
- This indicates how presumptive our model is about the shape of the data

Variance

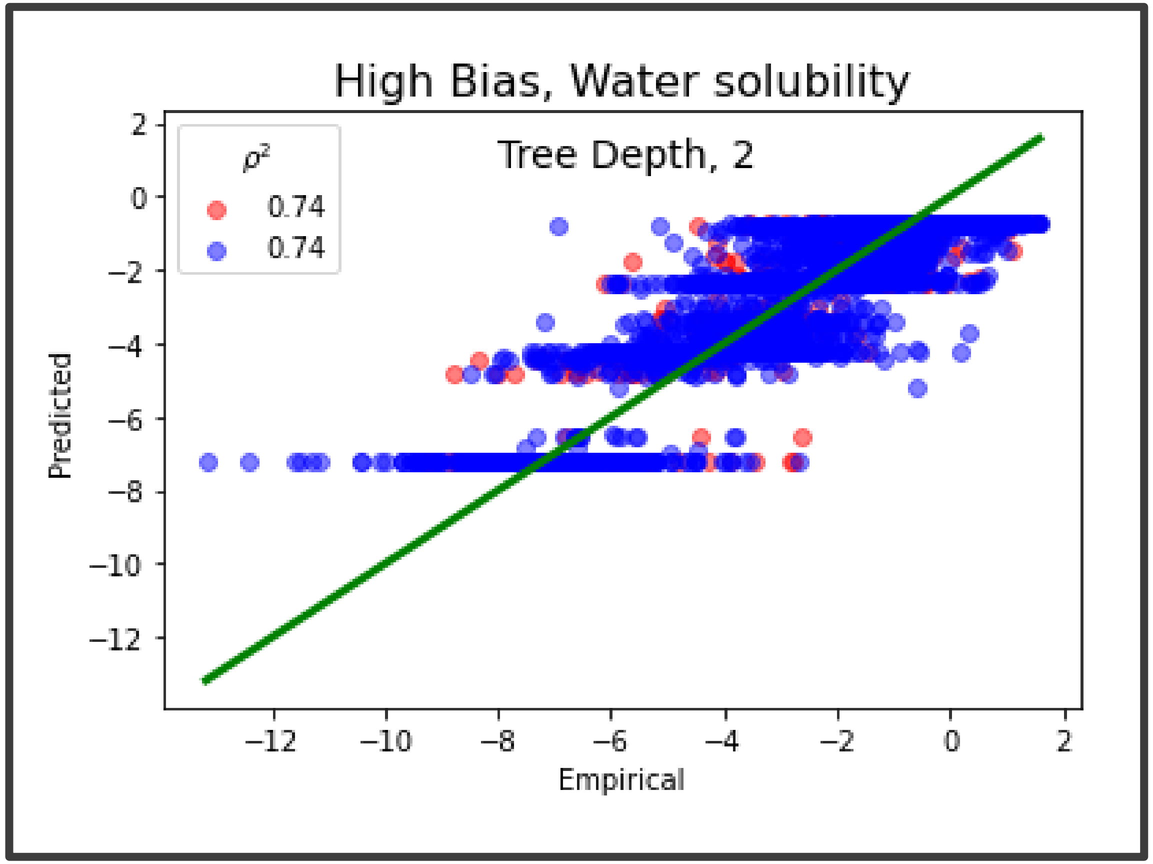
- This indicates how our model generalizes across the space of possible inputs

BIAS & VARIANCE

- We can think of these as characterizing of how much 'information' about the training set a model learns
  - High variance models learn a lot of information – possibly including noise or false signals!
  - High bias models learn relatively little information – you had better hope your model's fixed 'shape' is appropriate for the data!
  - Generally, decreasing bias increases variance and vice-versa
- The ideal model has <u>low</u> bias and <u>low</u> variance
- We control these through *hyperparameters*

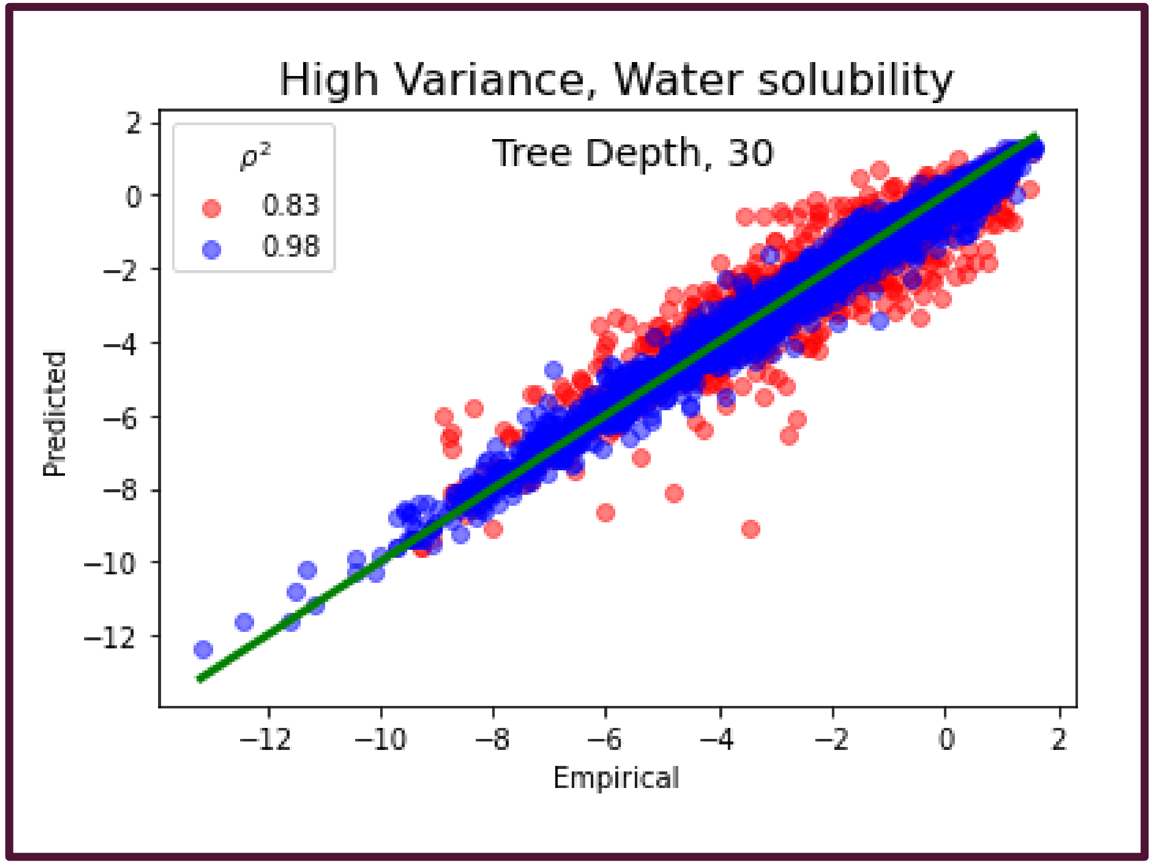


BIAS & VARIANCE – HIGH BIAS



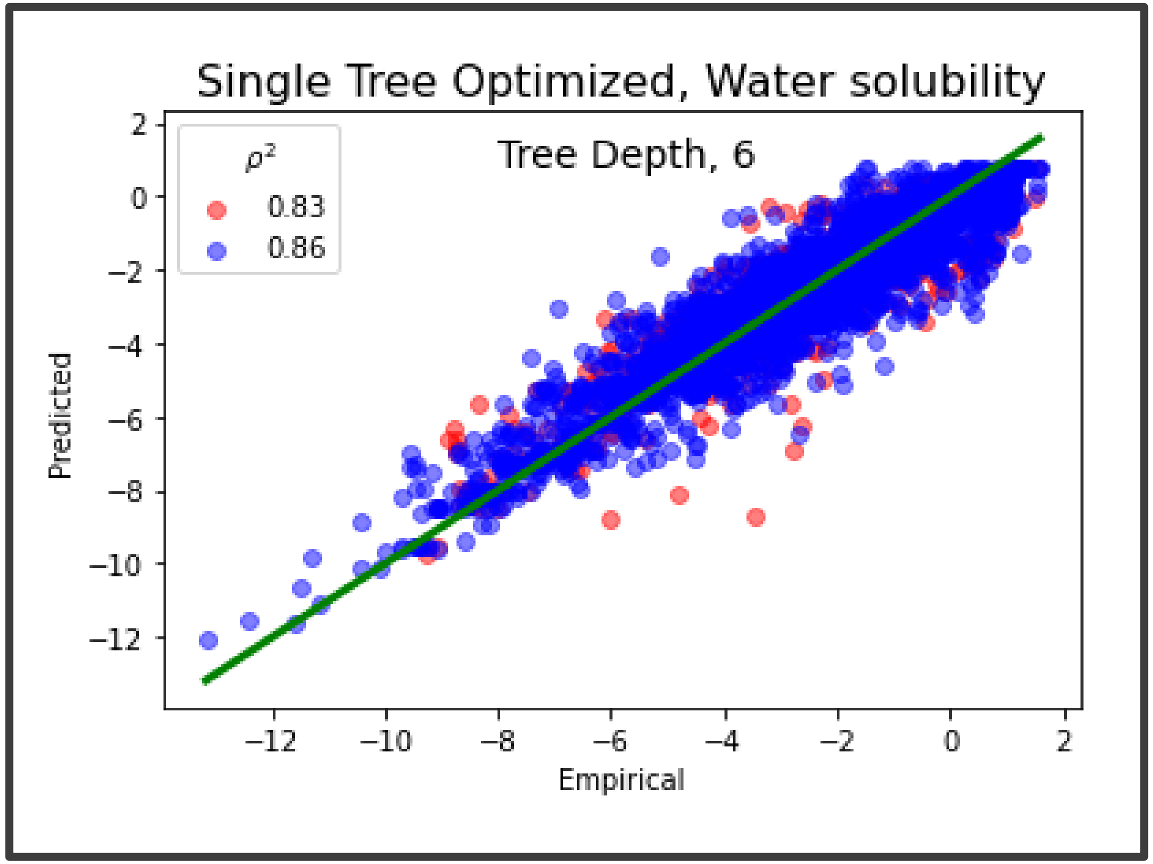
- Too Biased To See The Truth
  - We have 'assumed' too much by asserting our data can be described with decision trees of a relatively shallow depth of 2
  - The performance on both internal and external sets are relatively poor
  - The 'streaking' we see is due to the forest having too few terminal ensembles of points from which to derive an average

BIAS & VARIANCE – HIGH VARIANCE

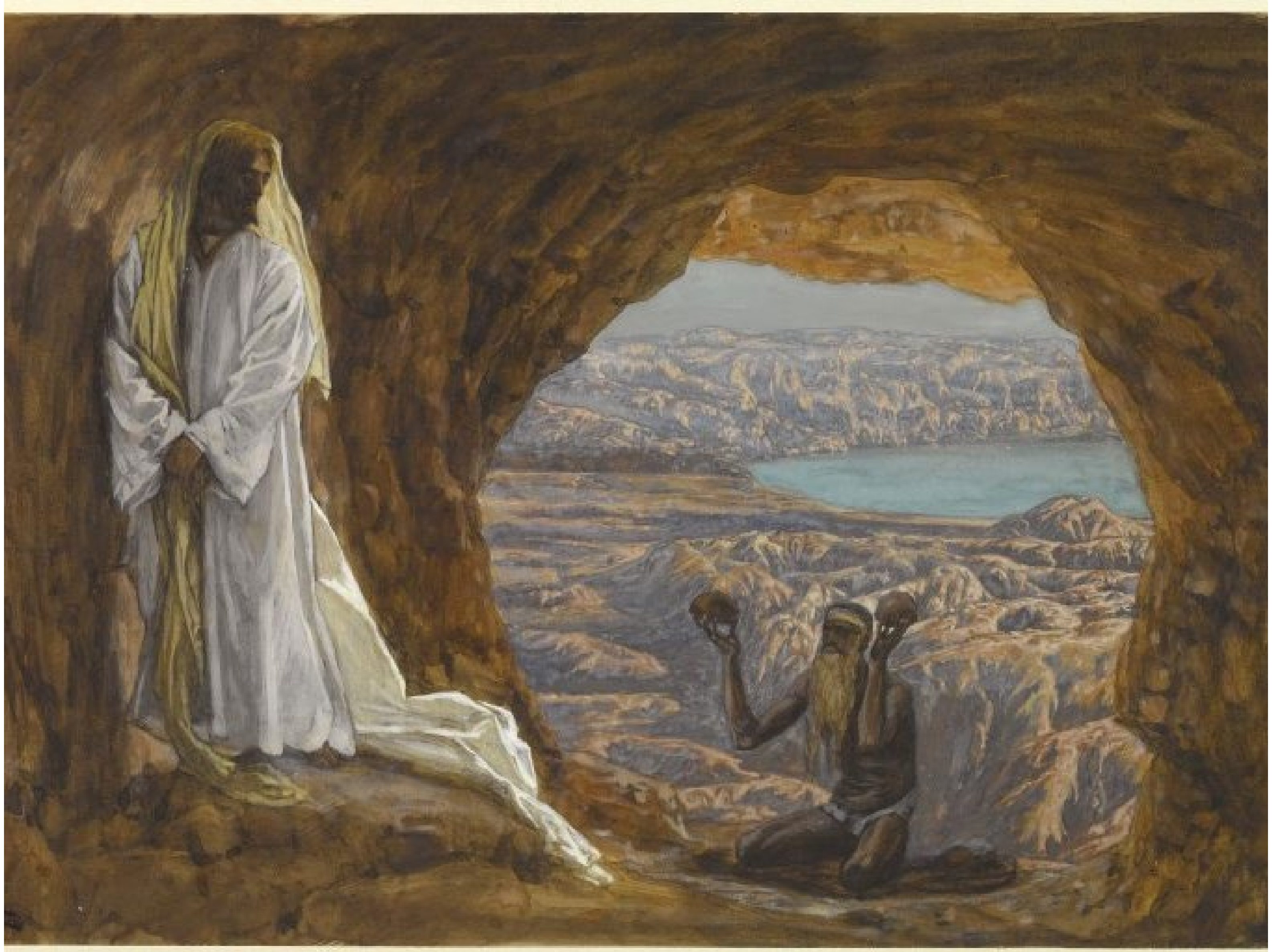


- A Manic Pixie Dream Model
  - Wow! These scores are so much better!
  - The behavior between the internal set (blue) and external set (red) is notably different
  - The red set was taken at random from the same region of chemical space as the blue set…what would happen if we go even slightly outside that region…?
  - Intimately tied to the concept of the "Applicability Domain"

BIAS & VARIANCE – THE STABLE SOLUTION

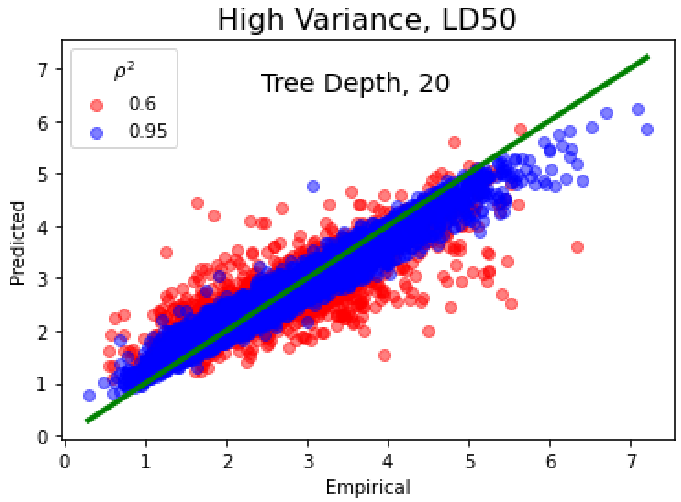


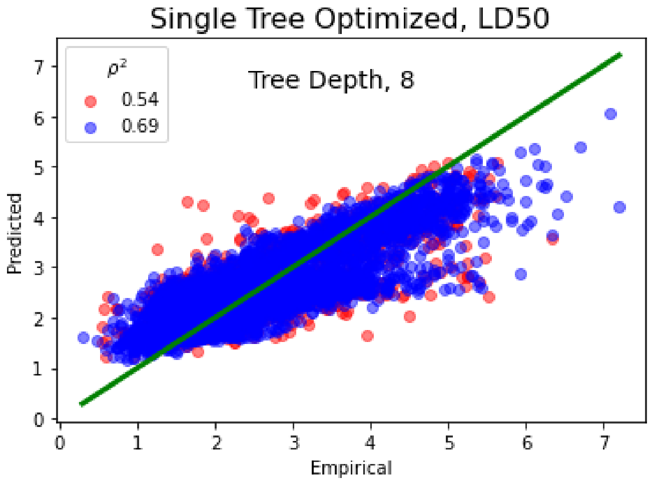
- Low Bias, Low Variance
  - The external score (red) is the same
  - Whatever our model learned, it seems to behave pretty similarly between the internal data and the external data
  - We allocated enough information capacity to pick up the major signals driving the data prediction *without* allocating so much we learned every quirk of the training set

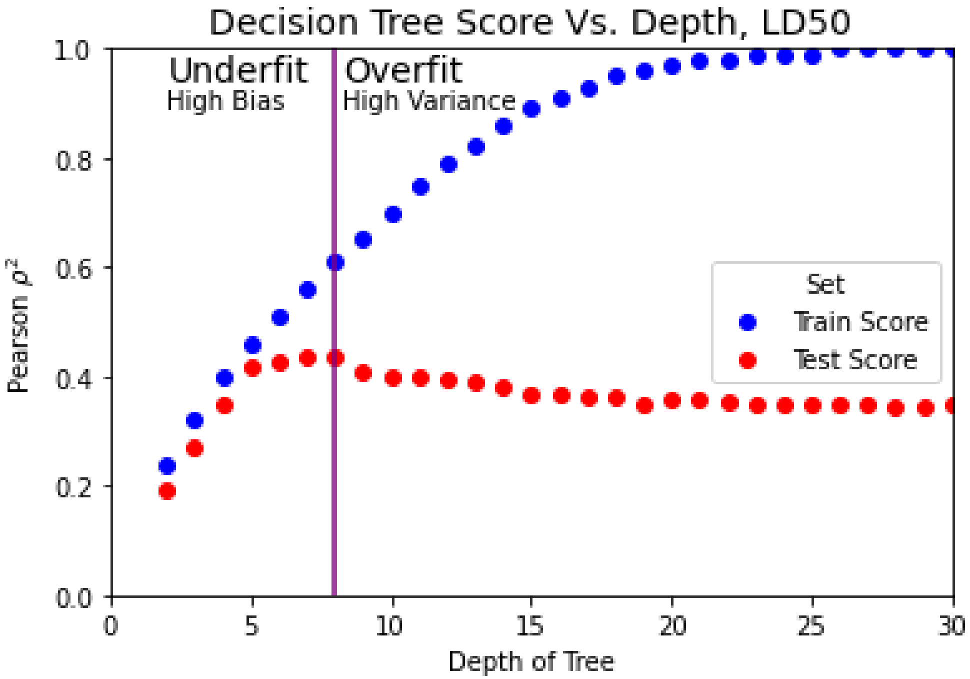


Jésus tenté dans le désert. James Tissot. {{PD-US-expired}} – published anywhere (or registered with the U.S. Copyright Office) before 1927 and public domain in the U.S.

IS HIGH VARIANCE WORTH IT FOR HIGHER EXTERNAL PERFORMANCE?

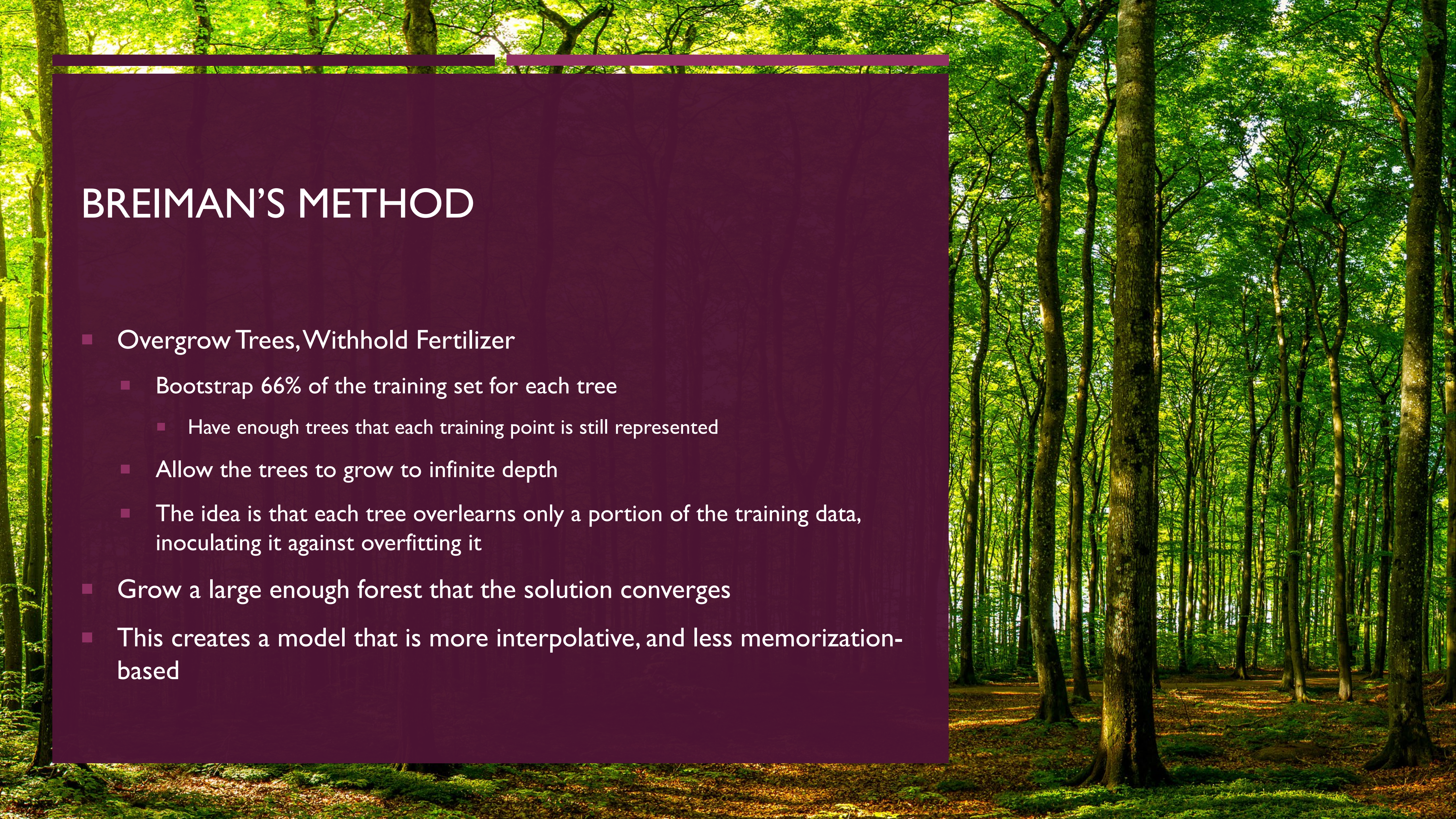






CAN AN ALGORITHM MAKE BREAD FROM STONE?

WHAT IF EXTERNAL PERFORMANCE LOOKS IMPROVED BY A GREATER DEGREE OF DIVERGENCE BETWEEN INTERNAL AND EXTERNAL SETS?



BREIMAN'S METHOD

- Overgrow Trees, Withhold Fertilizer
  - Bootstrap 66% of the training set for each tree
    - Have enough trees that each training point is still represented
  - Allow the trees to grow to infinite depth
  - The idea is that each tree overlearns only a portion of the training data, inoculating it against overfitting it
- Grow a large enough forest that the solution converges
- This creates a model that is more interpolative, and less memorization-based

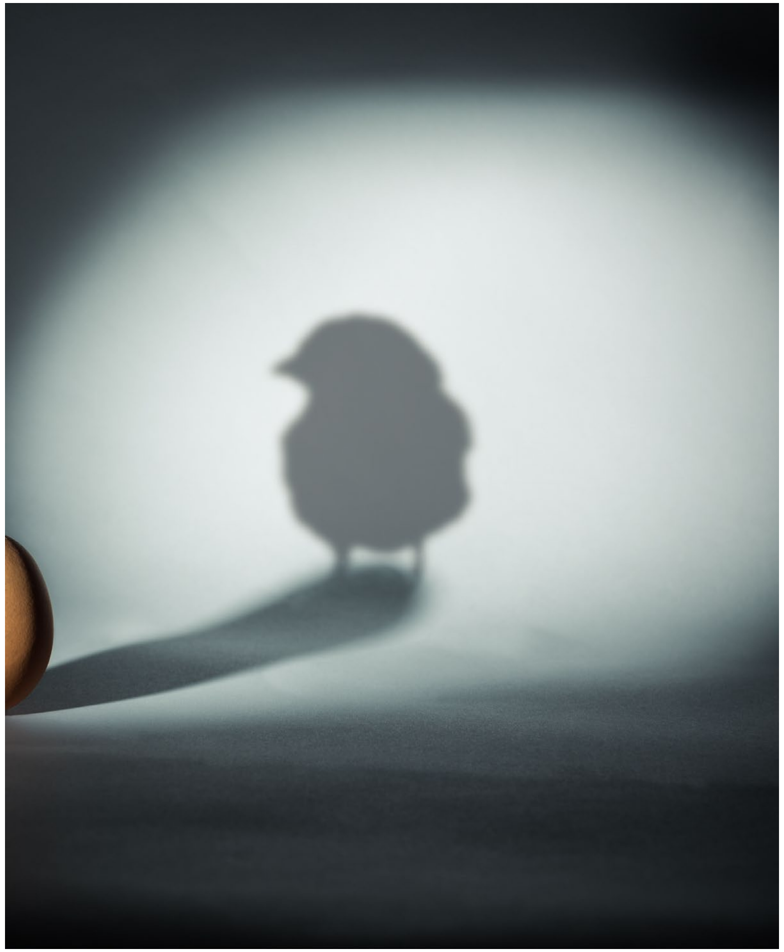
EMBEDDING CHEMICAL SPACE

"EMBEDDING" – A LOW DIMENSIONAL REPRESENTATION OF A HIGH DIMENSIONAL SPACE

By 4edges - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=73850232

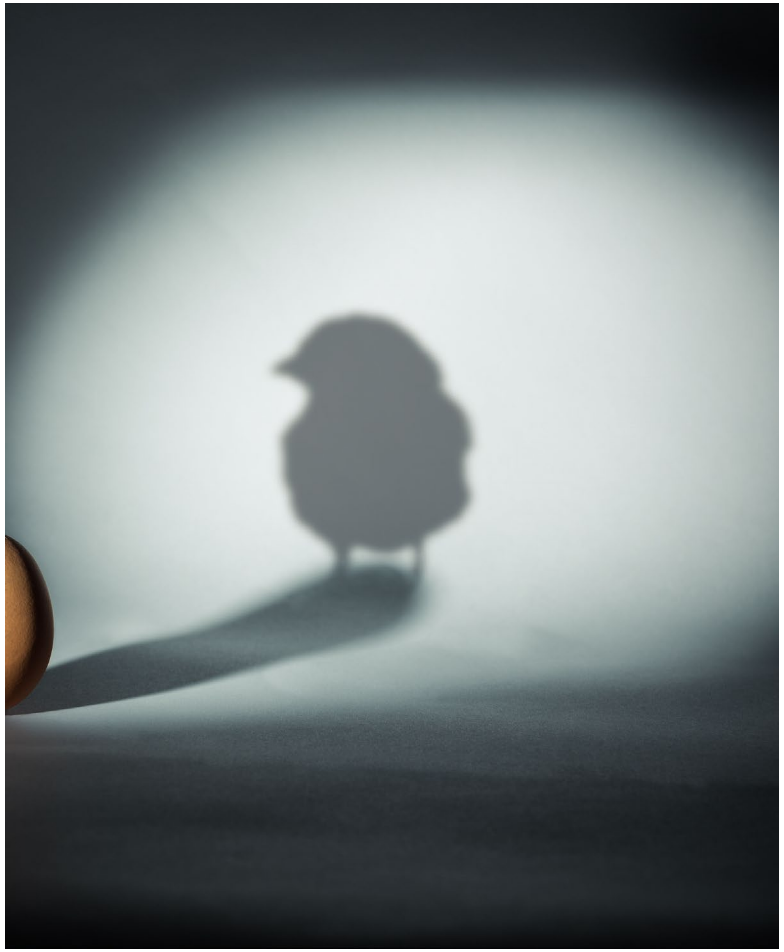
HOW DO WE BEST DESCRIBE CHEMICAL SPACE?

- What do we show a model as we try to teach it a QSAR?
  - We can show it as much as possible, but then we cede control of what it learns and how to interpret it
  - We can show it relatively few things, possibly compromising what it can detect but letting us more easily understand what it says
- This is our single greatest moment to enforce interpretability
  - A representation like ToxPrints or other fragment count representations have clear, interpretable chemical meanings, but this may limit the completeness of our description
  - A representation like T.E.S.T. or PaDEL is mathematically exhaustive in its description of structure, but can result in descriptors that are extremely abstract and difficult to relate to intuitive chemistries

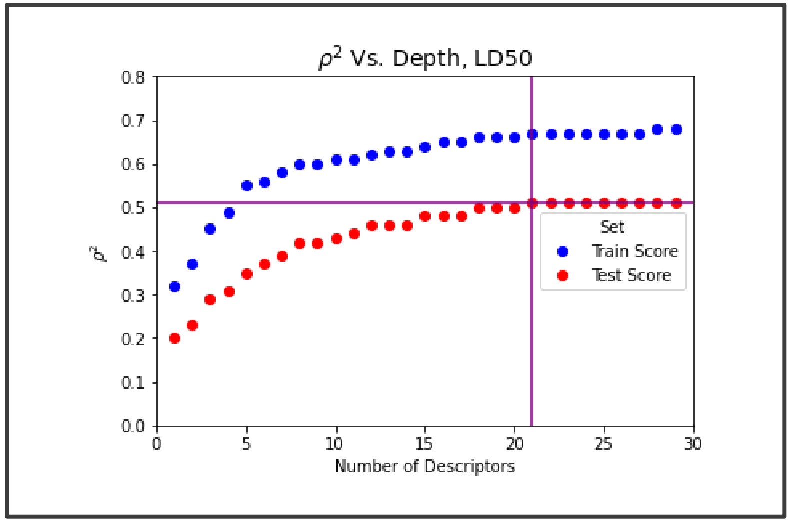


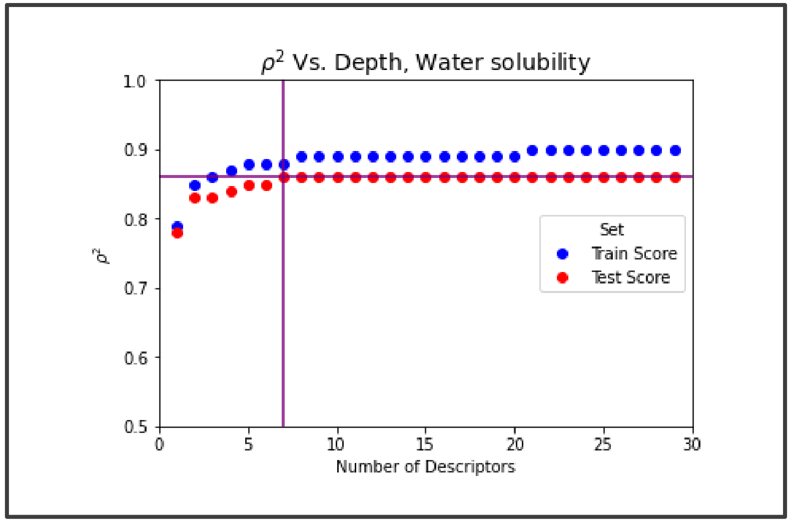
HOW DO WE BEST DESCRIBE CHEMICAL SPACE?

- "Show it *everything*"
  - PaDEL descriptors – 1875 Descriptors
  - T.E.S.T. descriptors – 979 Descriptors
- Advantages
  - Often higher performance statistics
  - If there's signal, that many descriptors will probably cover it
- Disadvantages
  - The curse of dimensionality
  - Translating the transmundane (it's harder to interpret)
  - If there's signal, that many descriptors will probably cover it



DO WE NEED ALL THOSE DESCRIPTORS?

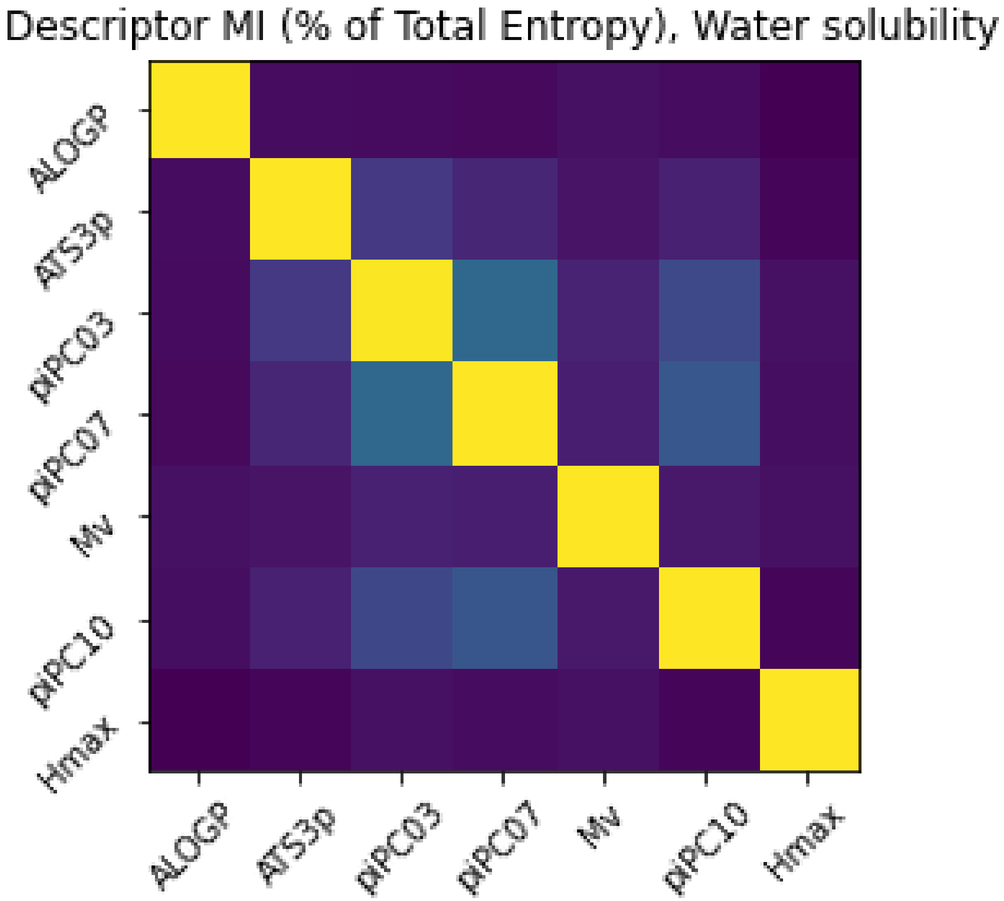




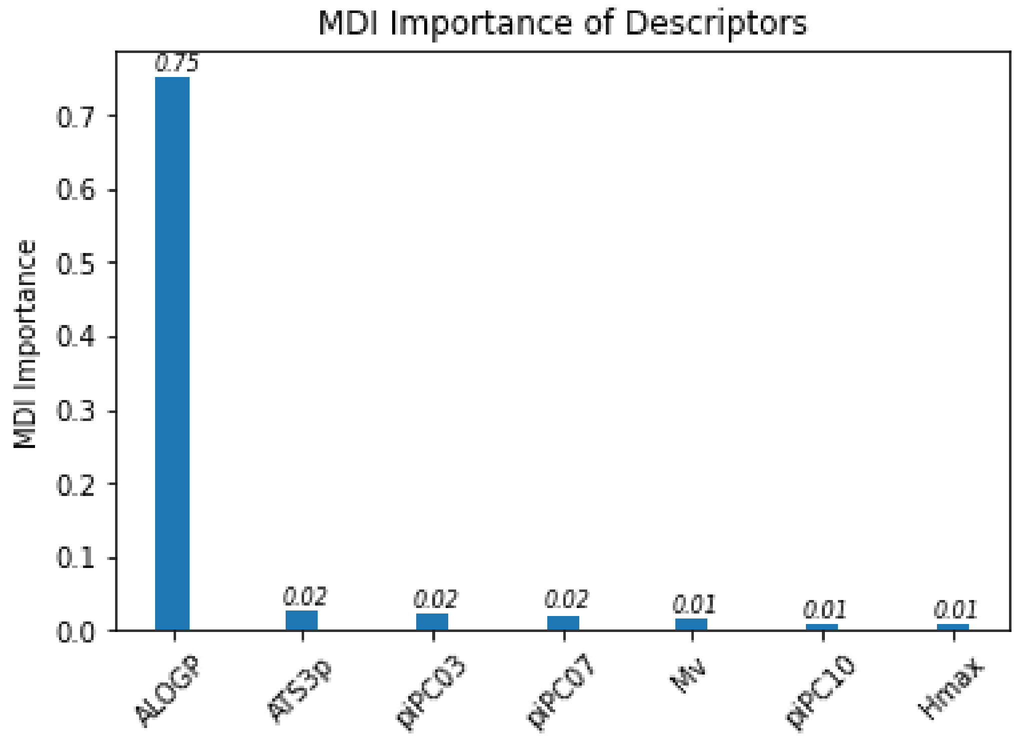
- Not Generally
  - More descriptors can increase overfitting due to more opportunities for erroneous patterns to emerge
  - Additional descriptors should add novel information, not repeat existing information
  - Mutual information is a robust but computationally expensive way to explore these relationships

DEVISING AN EMBEDDING

- Eliminate constant descriptors
  - They contain no information
- Eliminate highly colinear descriptors
  - These restate information
- Train a random forest on the remaining descriptors
  - Converge it properly, as the descriptors extracted from an overfit model do not necessarily indicate a general solution
- Pick the N most important descriptors, or the descriptors that are above some threshold fraction of the importance of the most important descriptor
  - Permutative Importance
  - Mean decrease in impurity
- Do the descriptors informatically overlap too much?

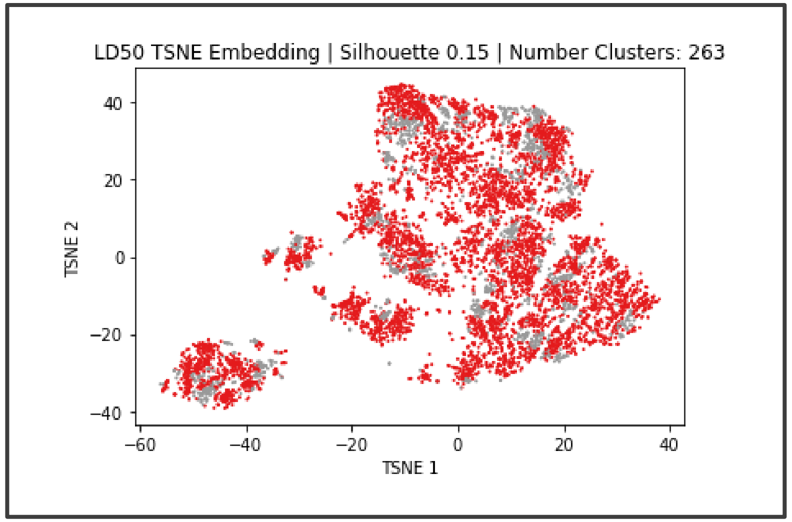


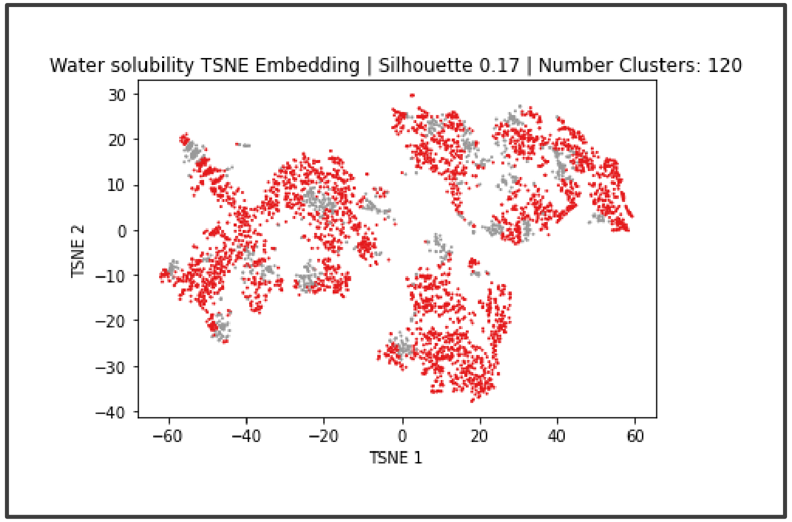
A POSTERIORI MECHANISMS



- Mechanisms for machine learning models are difficult to rigorously state
- The best process for reconciling mechanism with chemical intuition is analyzing the descriptors it utilizes
- Water solubility random forest model heavily weights ALOGP, the Ghose-Crippen octanol water coefficient, as the most important
  - This makes a lot of sense and can be considered an instance of transferring the information from the Ghose-Crippen model into the random forest model
- The remaining descriptors used are
  - Molecular multiple path counts – statements of bond order topology encoding information about aromaticity
  - Mean atomic van der Waals volume – affects exposed potential energy surface area defining interactions with solvent
  - Maximum hydrogen E-state value in molecule – rough statement of electronegativity possibility relating to polarity

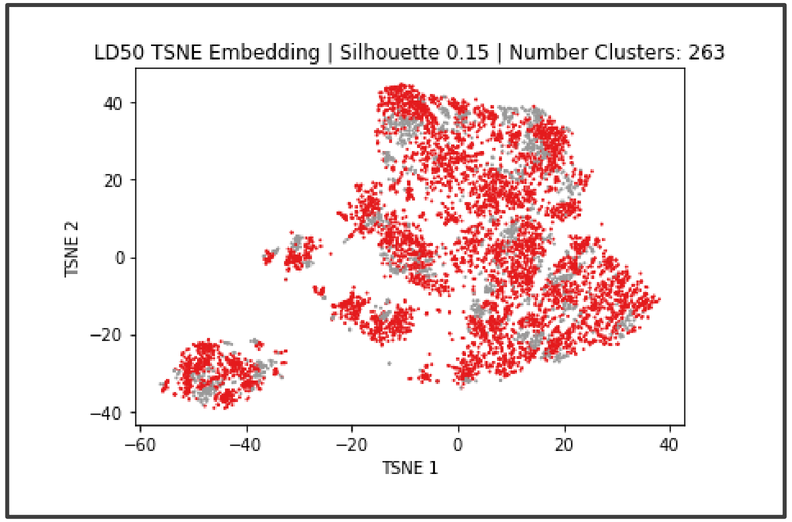
EXTERNAL TESTING

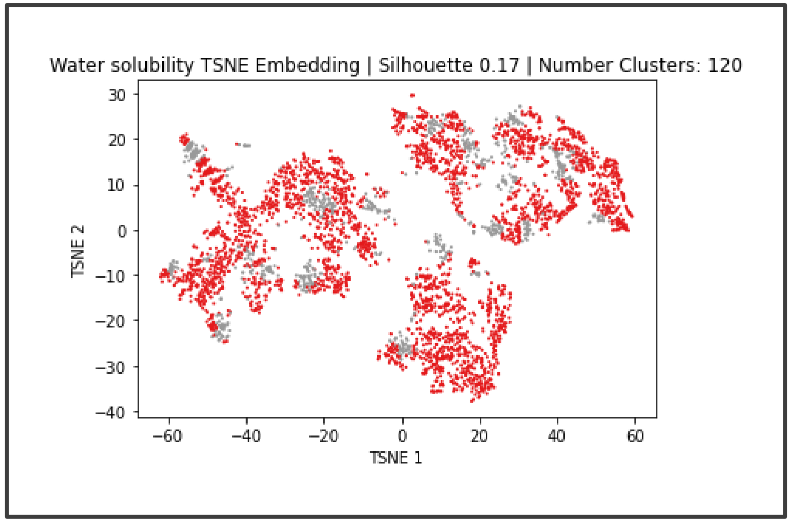




- Up until now, we've engaged in the common literature practice of using an external set randomly selected from the original data pool
- Can we rationally create a split that tests data external to the chemical space that was learned?
- Relatively novel territory

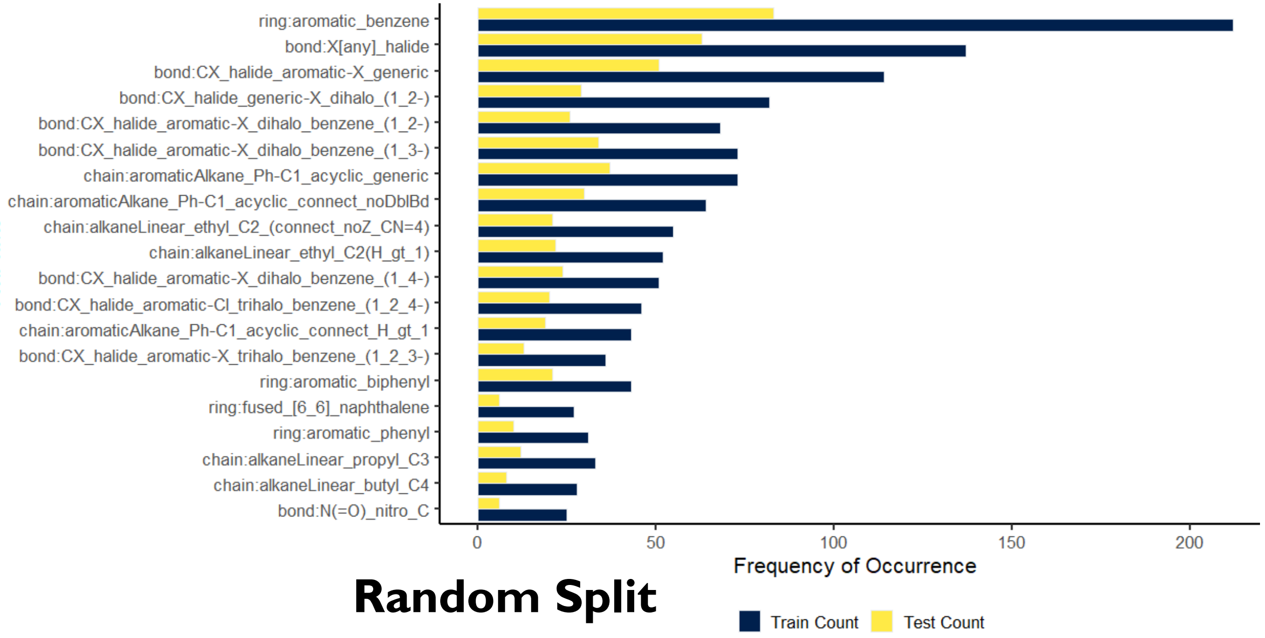
EXTERNAL TESTING



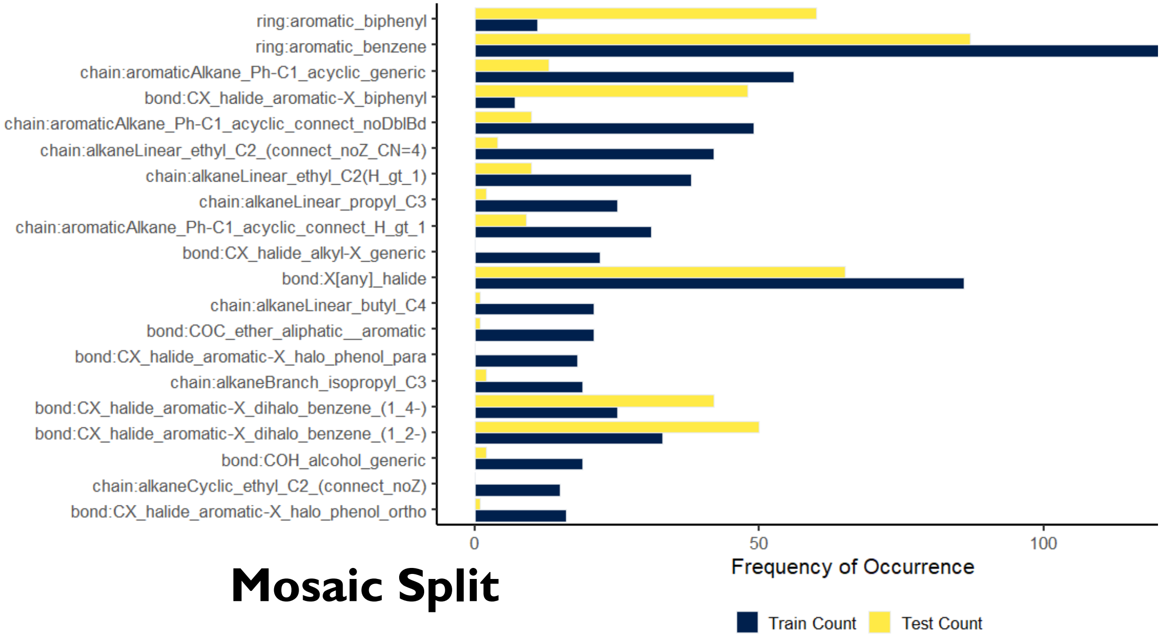


- The mosaic split uses affinity propagation clustering to isolate collections of samples in the embedded chemical space
- Those clusters can be recombined to create training and test sets
- This directly favors chemical dissimilarity between the sets

Figure credit: Dr. Charles Lowe

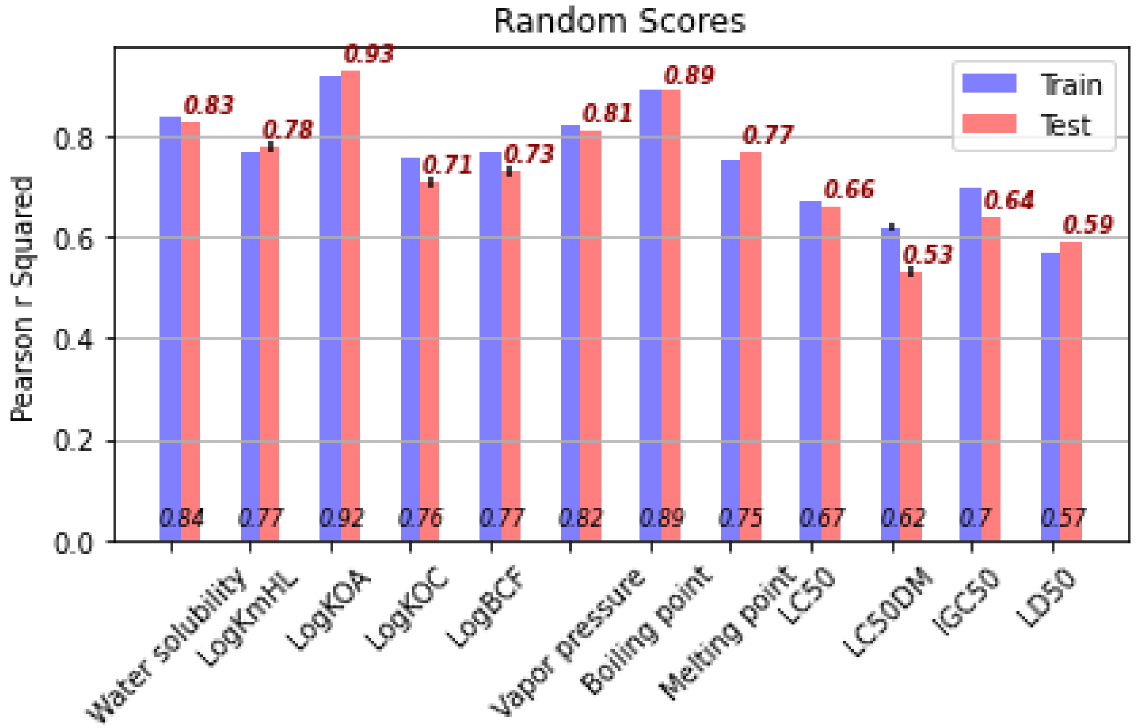


Random Split

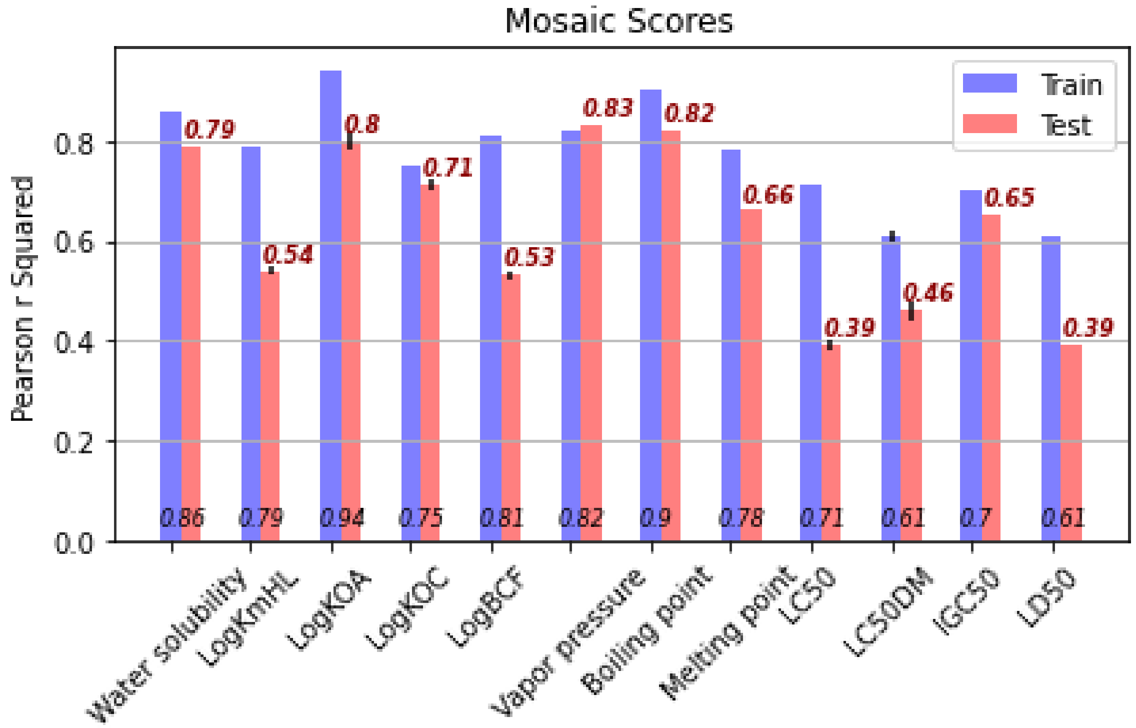


Mosaic Split

EXTERNAL TESTING



Random Scores



Mosaic Scores

EXTERNAL TESTING

REPORTING

- QMRF – QSAR Model Report Format
  - The gold standard of reporting information on a QSAR model
  - Sections are created around the OECD principles
- OECD Principles
  - 1. Define Endpoint
  - 2. Define Algorithm
  - 3. Define Applicability Domain
  - 4. Internal & External Validation
  - 5. Mechanistic Interpretation

THE MODEL PROCESS

- Putting It All Together
  - Select a structural representation that provides the desired balance of interpretability and completeness of description
  - Filter descriptors for redundant information and apply MDI selection to derive an embedding
  - Control the variance and bias through hyperparameters to converge a solution that performs similarly on internal and external data
  - A posteriori a mechanism based on the importance of the descriptors to the model
  - Run tests to consider external prediction ability of the model
  - Report the details in QMRF for publication

ACKNOWLEDGEMENTS

- Antony Williams & Todd Martin
- Gabriel Sinclair, Charles Lowe & Christian Ramsland
- CCCB
- Various Artists