Supervised Learning: Foundations for Chemical Data

CCE AI4ChemicalSciences Bootcamp

Dr. Sabrina Wahler

Models Covered

1. Basics
2. Linear Regression
3. Gaussian Process Regressor
4. Decision Tree Models
5. Summary

What do we need to do to make a model?



Choose Model



Hyperparameter tuning



Well curated database



Featurization



Computer doing Computer-Stuff



Model

Energetic Materials - Application

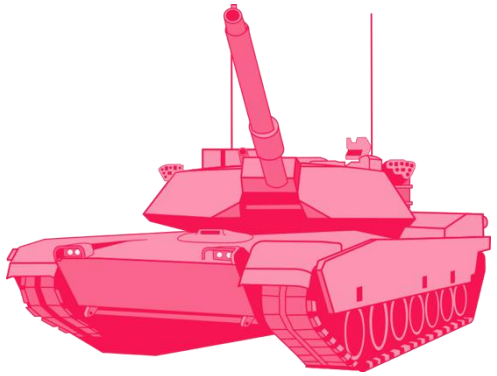
“Materials that chemically store a high amount of energy”



Aerospace



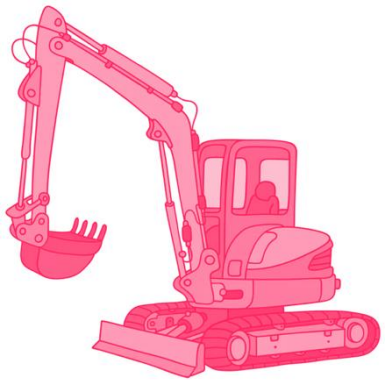
Mining



Defense



Pyrotechnics



Construction

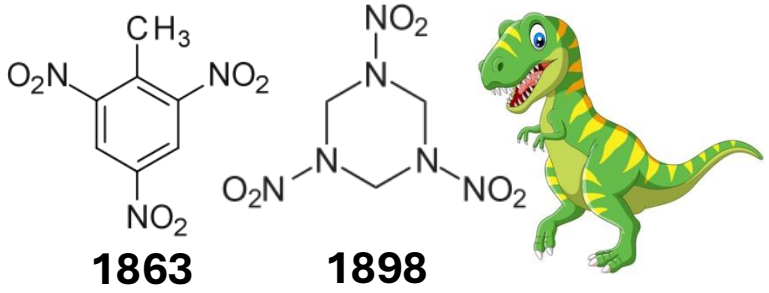
Motivation



Prevention of Accidents

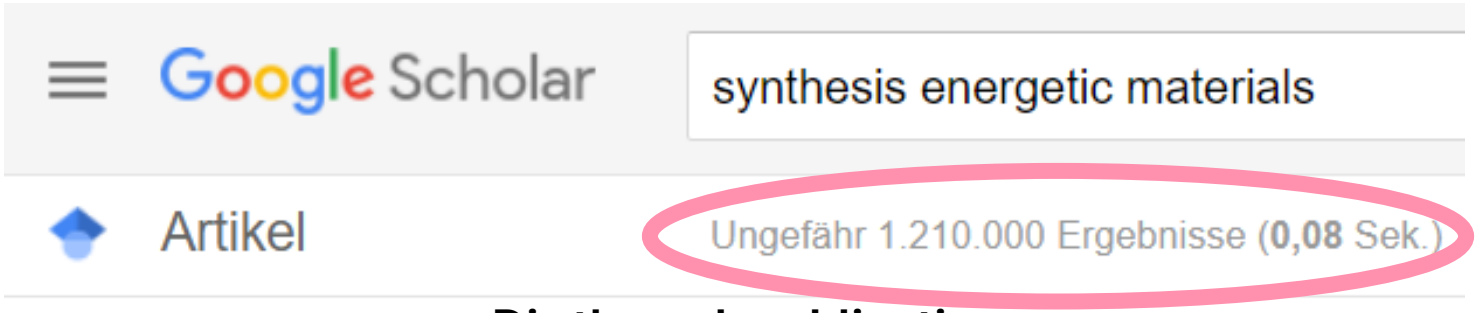


Reduce pollution



1863 **1898**

We still use ancient dangerous and toxic explosives today



Dig through publications

G. A. Giessler, A. Leopold, G. Germann, C. Heitmann, *Unfallchirurg* **2006**, *109*, 956–963.

What You'll Be Able to Do After Today

1. Definition of supervised learning
2. Explain why we split data into train / validation / test, and spot chemistry-specific pitfalls like scaffold leakage
3. Decompose prediction error into bias, variance, and noise — and read a learning curve
4. Tell model parameters apart from hyperparameters, and tune the latter without contaminating the test set

Along the way: a full worked case study, a real example of data bias becoming model bias, missing data in a chemistry context, and three baseline models worth always trying.

OBJECTIVE 1

What Is Supervised Learning?

Learn a function **f(X) → Y** from labeled examples — pairs of input X and correct output Y — so it can predict Y for a new, unseen X.

Two flavors:

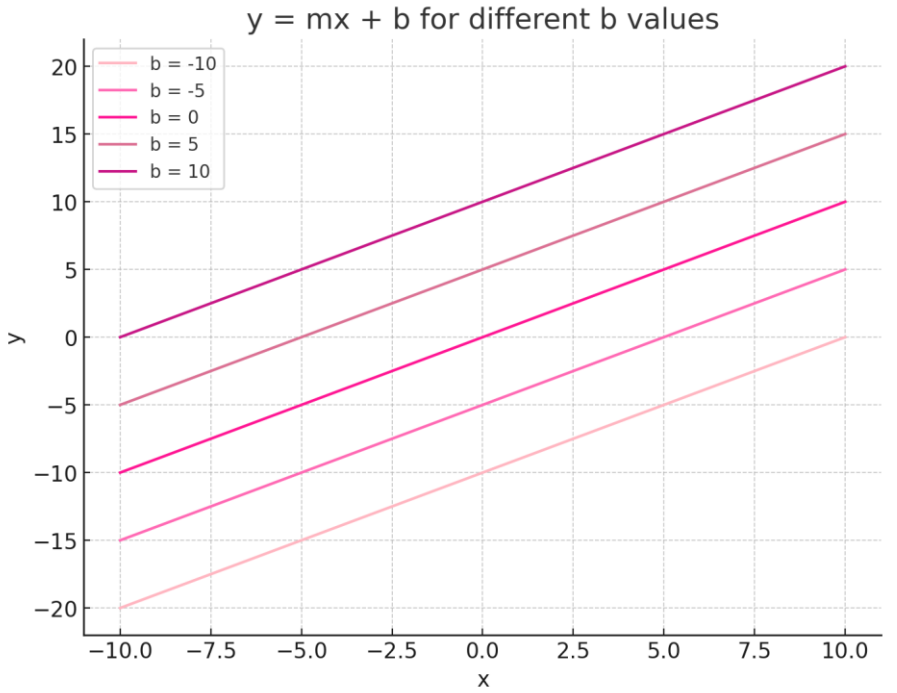
Regression
continuous output (e.g. a yield, a velocity, a temperature)

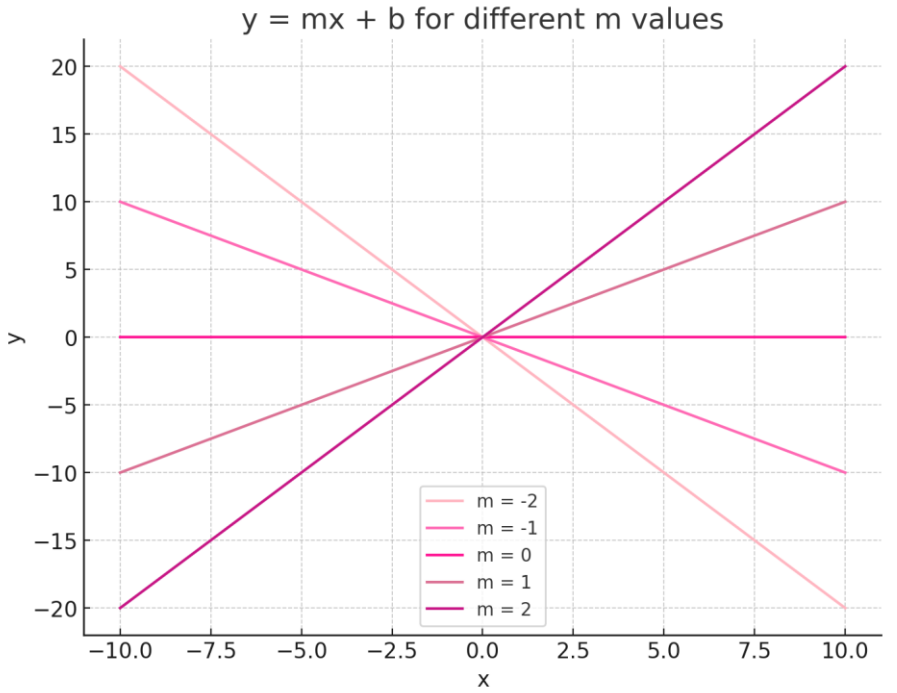
Classification
discrete label (e.g. toxic / non-toxic, sensitive / insensitive)

Not covered today: unsupervised learning (no labels — finds structure in data) and self-supervised learning. GNNs, VAEs, and Transformers later in the bootcamp build on the same fit-to-labels idea with richer architectures.

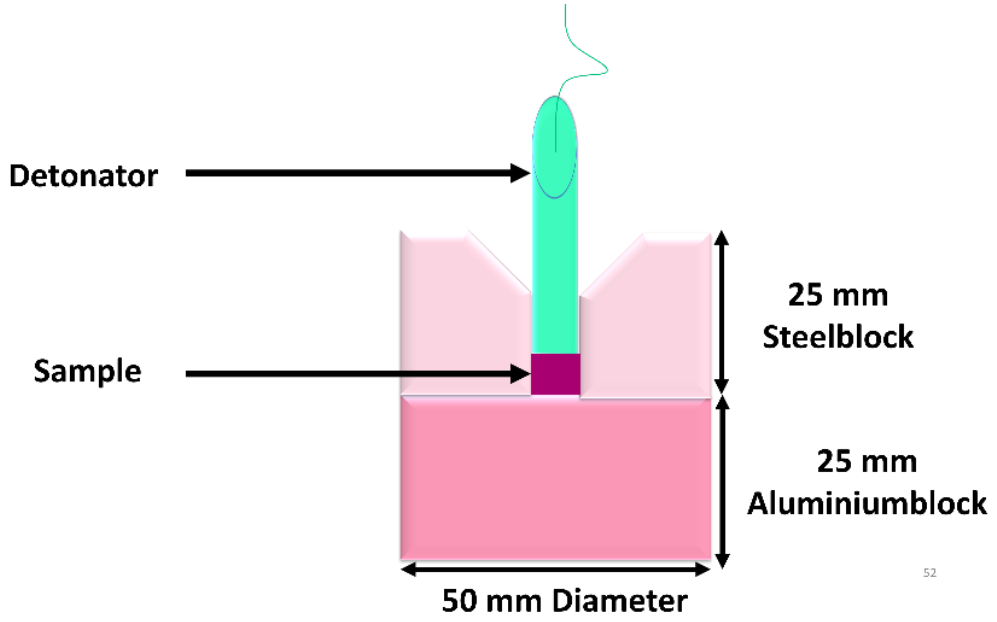
Linear Regression – The simplest form of ML…

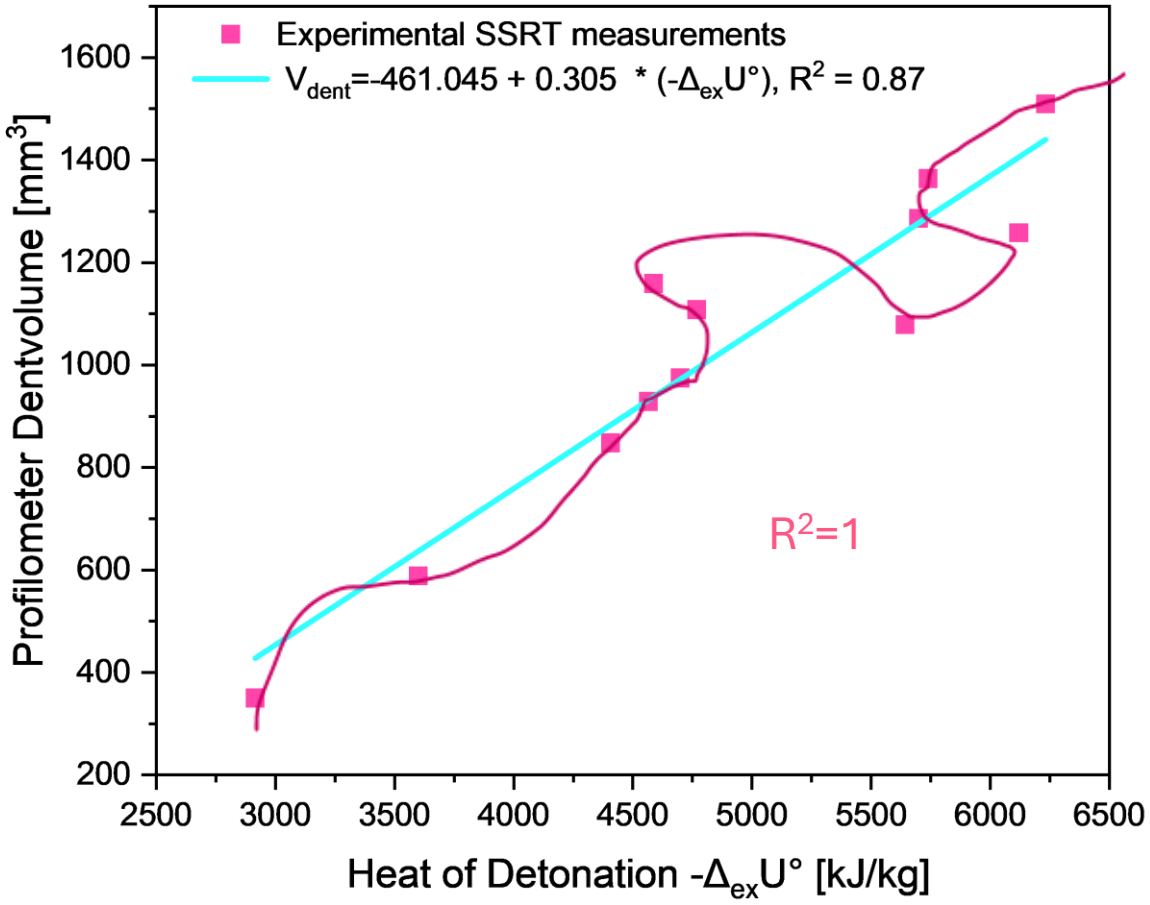
- Simplest form of Machine Learning y = mx + b
- Allows for simple interpretation, learning from the labels X and Y
- May miss complex relationships





SSRT Dent Test

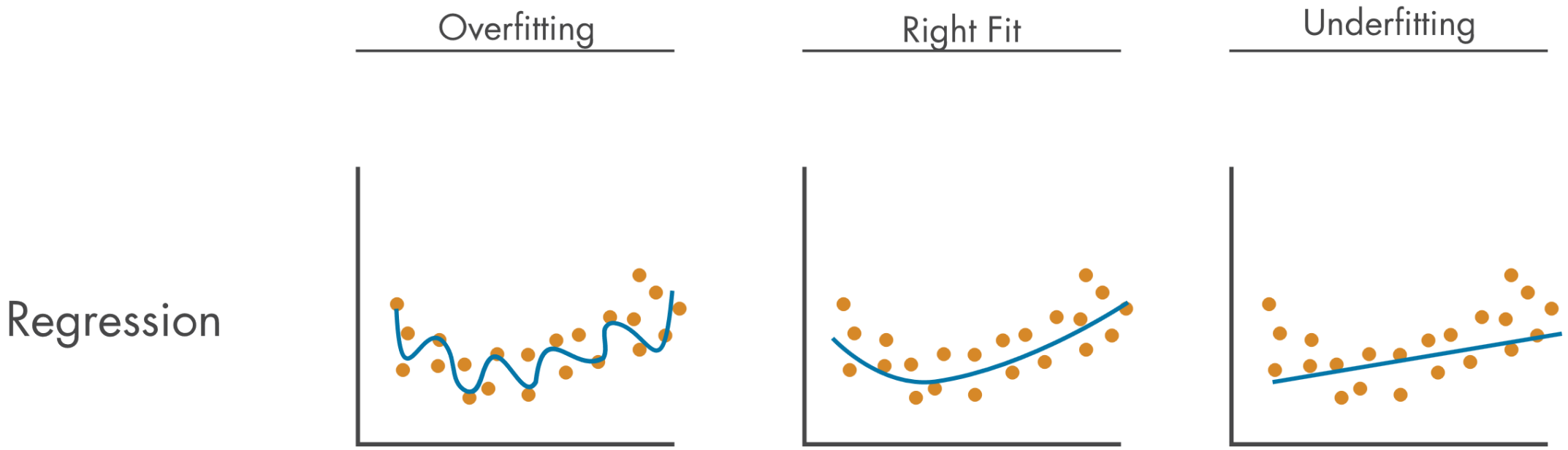




SSRT-Dent [mm³]= = (-461.045) + (0.304992) * (- $\Delta_{ex}U^{\circ}$[kJ/mol])

Wahler, S., Klapötke, T. M., Int. J. Energ. Mater. Chem. Propuls. **2023**, Accepted Manuscript

Underfitting vs. Overfitting in a Regression



Underfit = high bias. **Overfit** = high variance. Goal: the sweet spot between them.

Metrics to evaluate Model Performance

Metric	Formula	Measures	Interpretation
MAE (Mean Absolute Error)	$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$	y_i - $\hat{y}_i$	
MSE (Mean Squared Error)	$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	Squared deviations from truth	Penalizes larger errors more heavily
RMSE (Root Mean Squared Error)	$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$	Root of squared error	Same units as the target; still sensitive to outliers
MAPE (Mean absolute percentage error)	$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left\lvert\frac{y_i - \hat{y}_i}{y_i}\right\rvert \times 100\%$	Relative error in %	Easy to interpret; shows average % error
R^2 Score (Coefficient of Determination)	$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$	Proportion of variance explained	1 = perfect, 0 = mean prediction, < 0 = worse than mean

OBJECTIVE 2

Why Do We Split The Data?

- A model can reach near-perfect performance on the data it was trained on — that tells you nothing about how it behaves on a new molecule.
- The only honest way to estimate real-world performance is to evaluate on data the model never saw during training.
- This is the difference between memorizing and learning.

Train / Validation / Test: Three Different Jobs



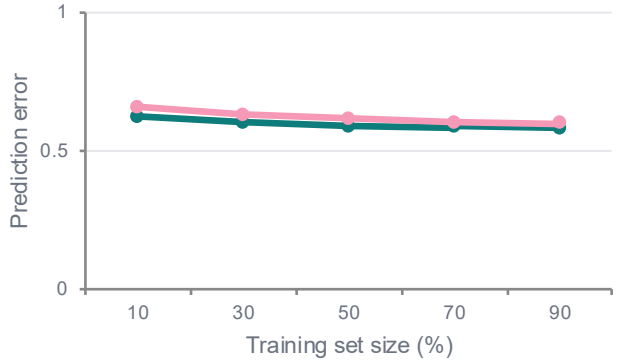
Test is touched exactly once — at the very end — for an honest final number.

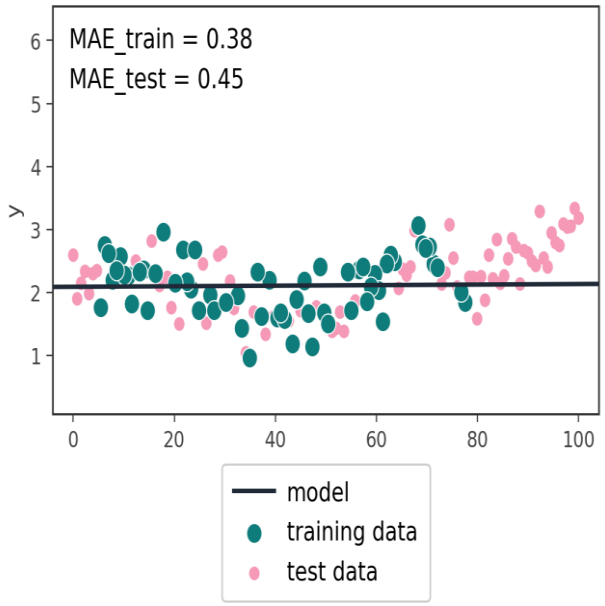
Diagnosing Model Fit with Learning Curves

Training vs. validation error as a function of training set size

1. Underfitting

High bias — both errors high & close

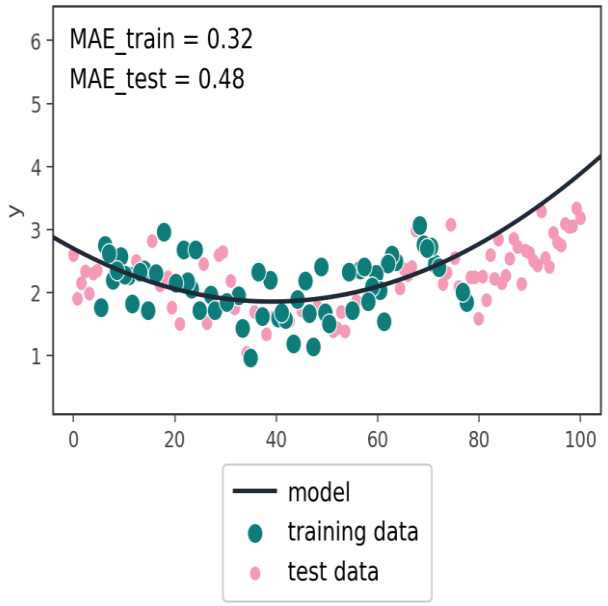




2. Good Fit

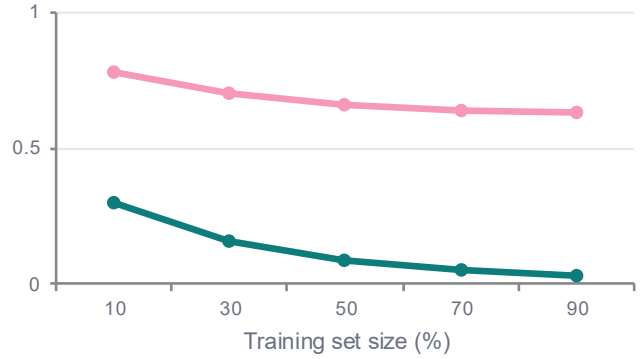
Errors converge to a low, stable value

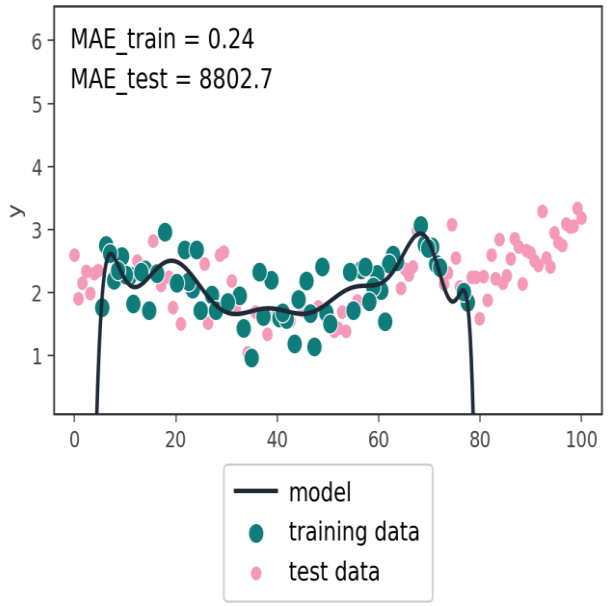




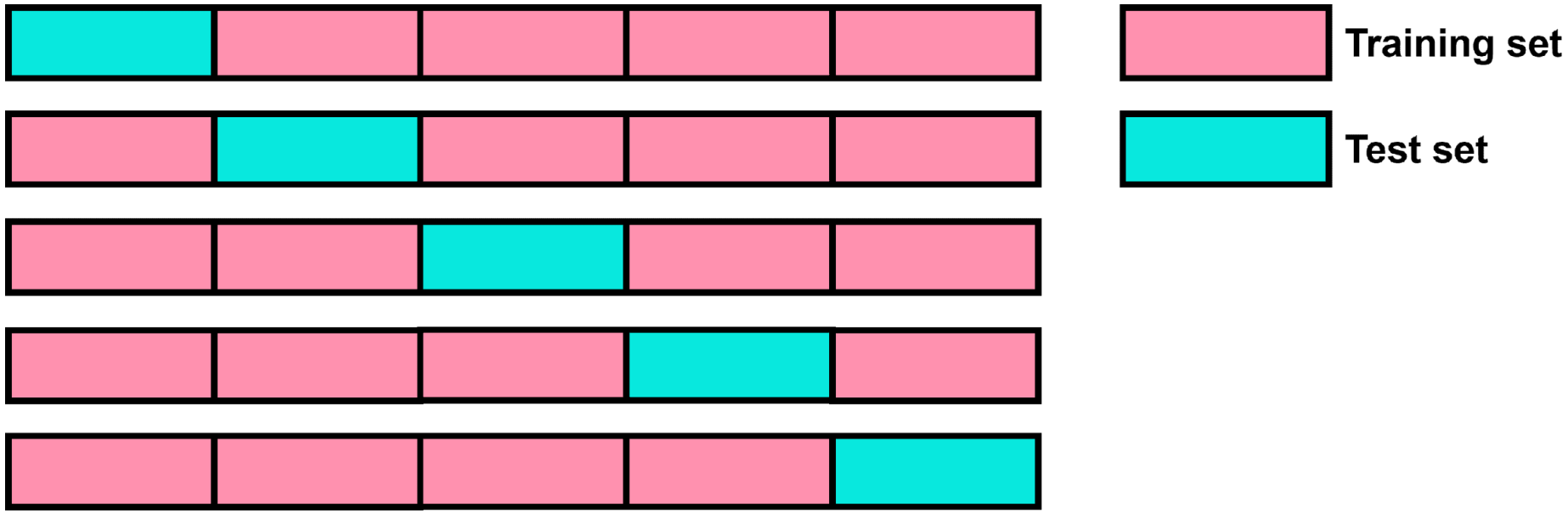
3. Overfitting

High variance — persistent large gap





K-Fold Validation



Monte Carlo

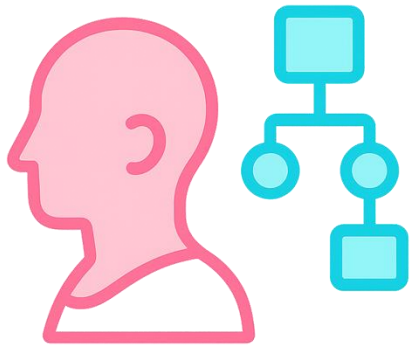


The Chemistry Pitfall: Why Random Splits Lie

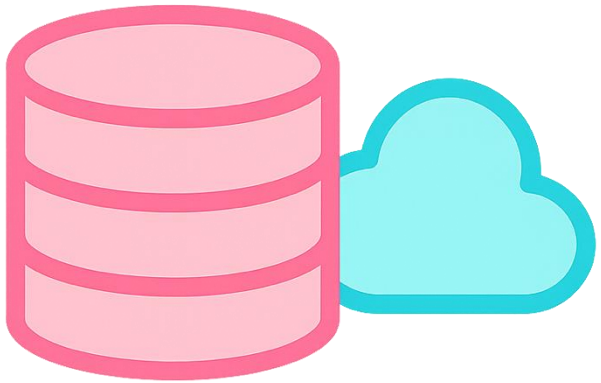
- Molecules cluster into structural families (scaffolds). Random splitting puts near-duplicates of the same scaffold on both sides.
- The model is effectively tested on molecules it has already seen → performance looks better than it really is.
- Fix: split by scaffold (e.g. Murcko scaffold) so an entire structural family is held out together.



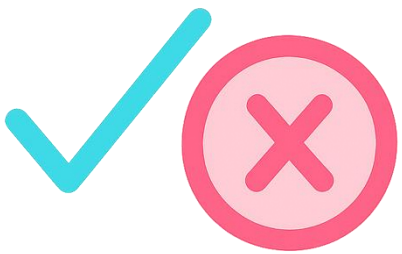
Avoiding overfitting



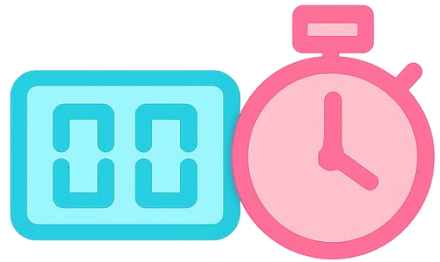
Simpler Model



More Data

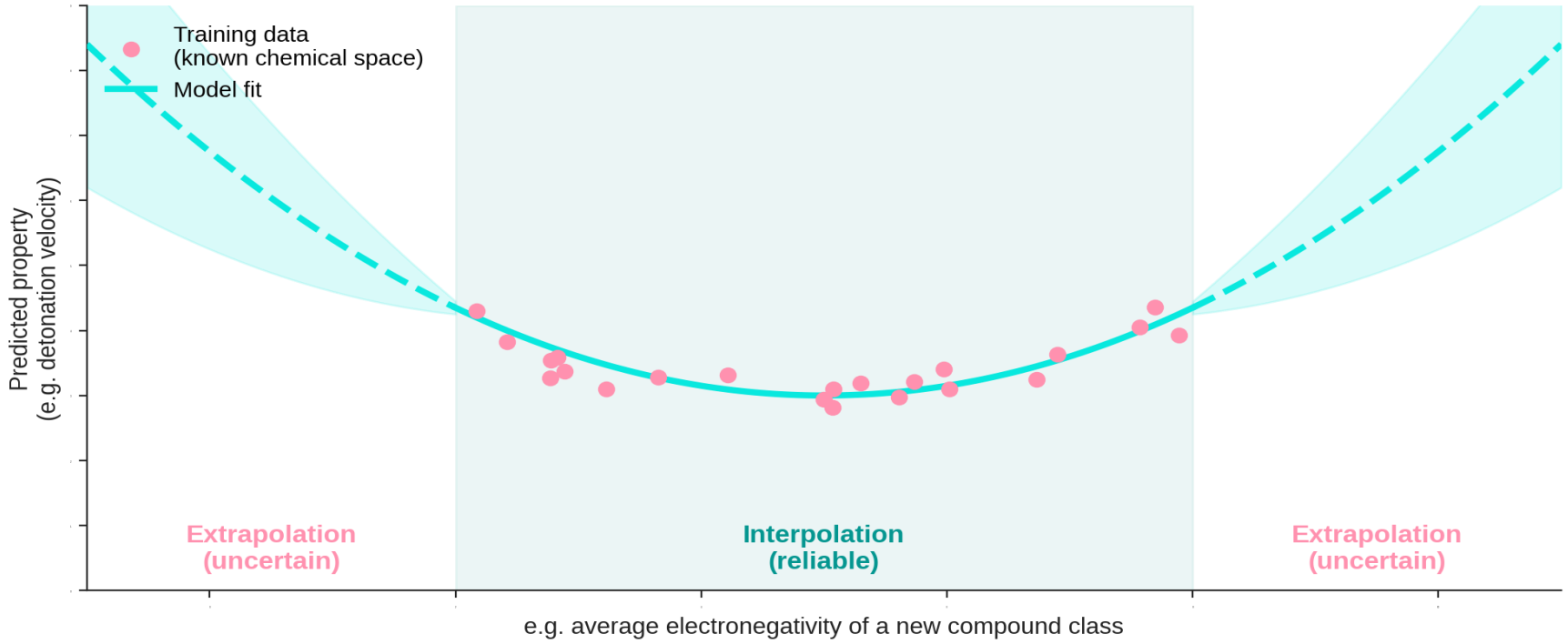


Cross Validation



Early Stopping

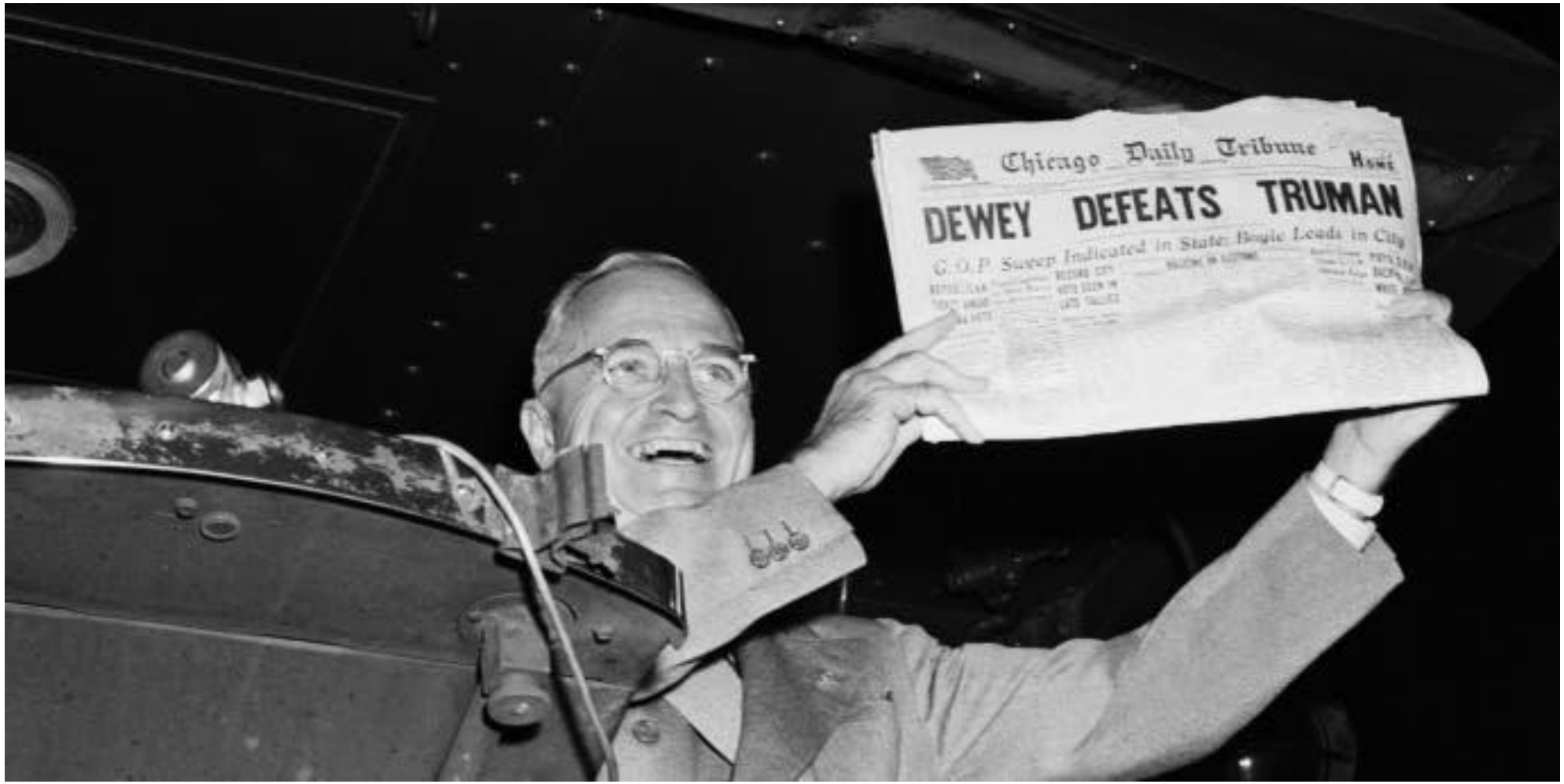
Interpolation Is Easy. Extrapolation Is Hard.



When Models Learn the Wrong Thing: Data Bias

- Models don't learn “chemistry” — they learn whatever pattern is easiest to fit in the data you gave them.
- If the data-collection process itself is biased, a model can hit great metrics by learning the bias instead of the underlying science.
- This is dangerous precisely because it's invisible in a standard train/test split — the bias is present on both sides.

Biases in collecting data



OBJECTIVE 3

Decomposing Prediction Error

$$\text{Expected Error} = \text{Bias}^2 + \text{Variance} + \text{Irreducible Noise}$$

Bias — Systematic error from a model too simple to capture the real relationship (underfitting)

Variance — Error from a model too sensitive to the specific training sample (overfitting)

Noise — Measurement error / randomness no model can remove (e.g. replicate-to-replicate variability)

Bias & Variance, Visually



What do we need to do to make a model?



Choose Model



Hyperparameter tuning



Well curated database



Featurization



Computer doing Computer-Stuff



Model

OBJECTIVE 4

Model Parameters vs. Hyperparameters

Model	Parameters (learned from data)	Hyperparameters (set before training)
Ridge / MLR	Coefficients β and intercept	Regularization strength λ
Random Forest	Split rules & leaf values in each tree	Number of trees, max depth
GPR	Posterior mean/covariance at the data points	Kernel lengthscale, noise variance

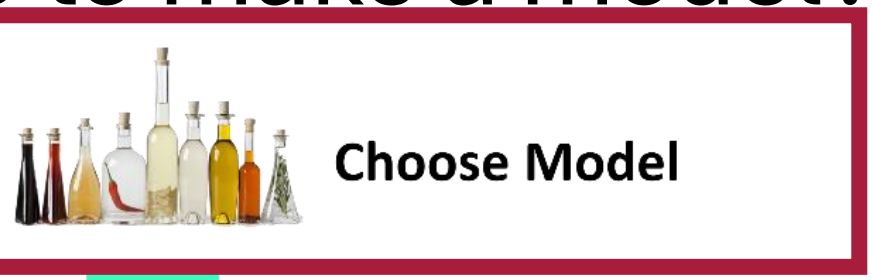
Tuning Hyperparameters Without Cheating

- Never choose hyperparameters by looking at test-set performance — that leaks test information into your model choice, and your final number will be optimistic
- Use the validation set (or cross-validation) to compare hyperparameter settings
- Only touch the test set once, at the very end, with the winning configuration

How do we fit?

Method	Description	Pros	Cons
Manual Search	Manually select hyperparameter values based on intuition.	Simple, fast for small problems.	Not systematic, may miss good values.
Grid Search	Exhaustive search over all parameter combinations.	Systematic, easy to parallelize.	Exponential growth with many parameters.
Random Search	Randomly sample parameter combinations.	More efficient than grid for many parameters.	Still inefficient for huge spaces.
Bayesian Optimization	Build a model of the objective function to guide search.	Efficient, balances exploration/exploitation.	More complex to implement.

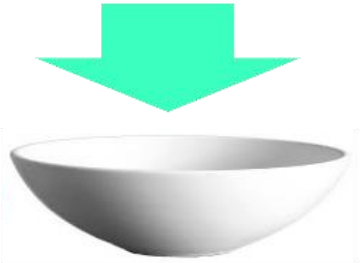
What do we need to do to make a model?






Well curated database

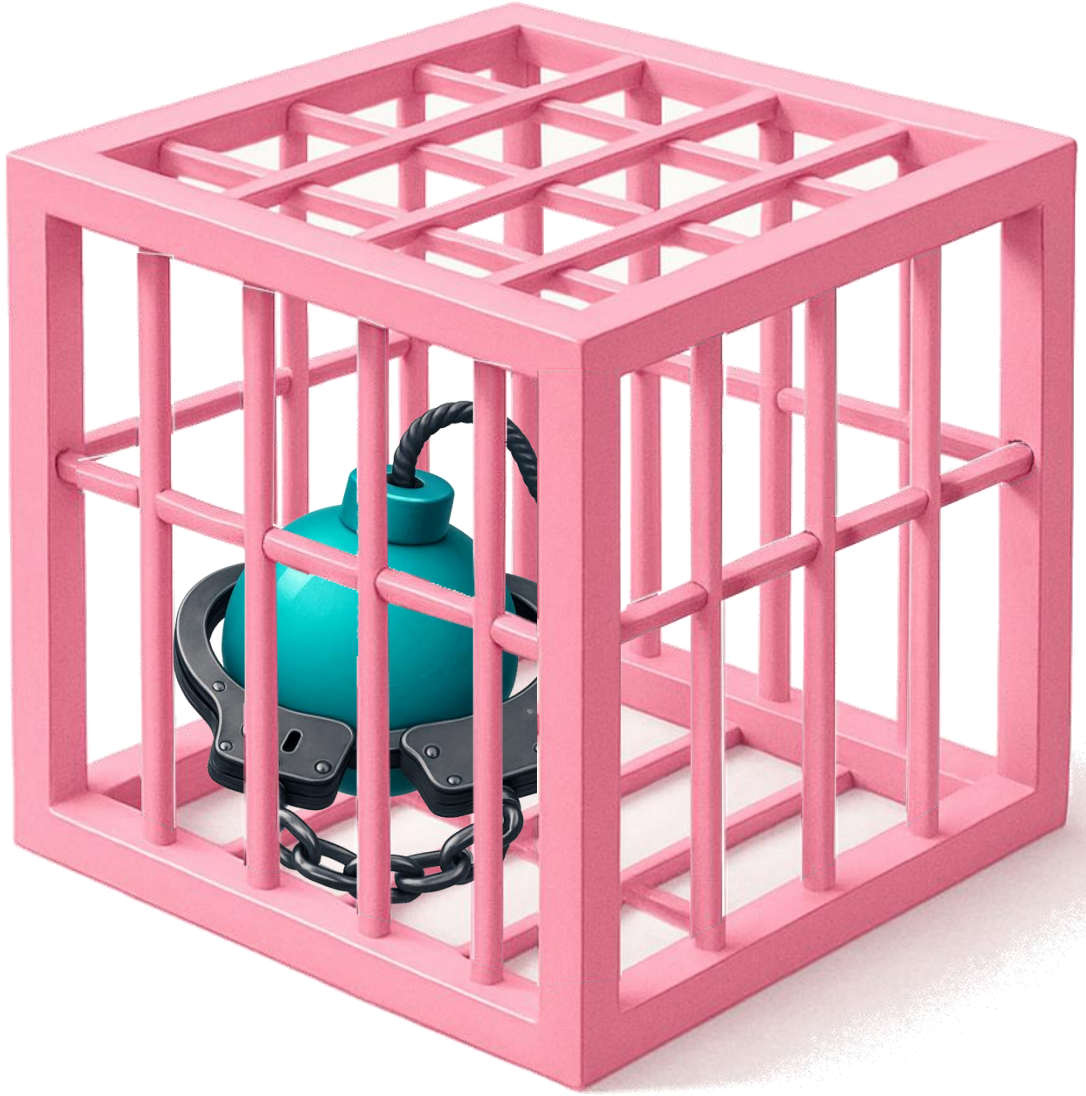

Featurization


Computer doing Computer-Stuff


Model

Current detonation models

Require Detonation Products and/ or elemental composition as input…



Propellants, Explosives, Pyrotechnics

WILEY-VCH

Propellants, Explosives, Pyrotechnics

RESEARCH ARTICLE OPEN ACCESS

Putting an End to the Handcuffs of the Periodic Table: Using Atomic Features for Predicting Detonation Velocity

Dedicated to Professor William A. Goddard III on the occasion of his 88th birthday.

Sabrina Wahler[1,2]

[1]Department of Engineering, Imperial College London, Institute for Security Science and Technology, London, UK | [2]Founder, RoseExplosive GmbH, Munich, Bavaria, Germany

Correspondence: Sabrina Wahler (sabrina.wahler@roseexplosive.com)

Funding: This work was awarded the BFSZ-Seal by the German Federal Ministry of Education and Research (Project ID: 017-414-775/2024-2).

ABSTRACT

In this study, a Gaussian process regressor is used to predict the detonation velocity. As features atomic values were used, which allows the application of the model to explosives of any elemental compositions. This approach does not require reparameterization

Model

Heat of formation and Electonegativty limited the interpolation due to smaller range in CHNO Explosives

RoseAtoms
Model 2

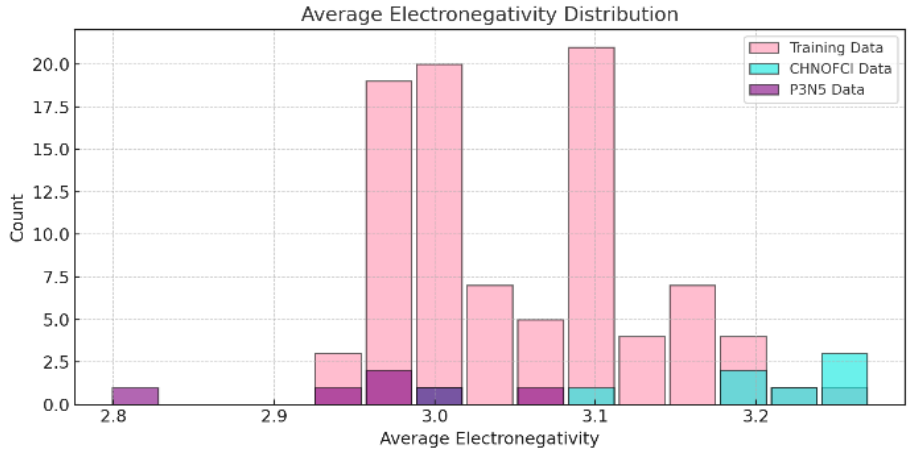
Number of Atoms
Number of Protons
Number of Neutrons
Number of Single Bonds
Number of Double Bonds
Number of Aromatic Bonds
Density

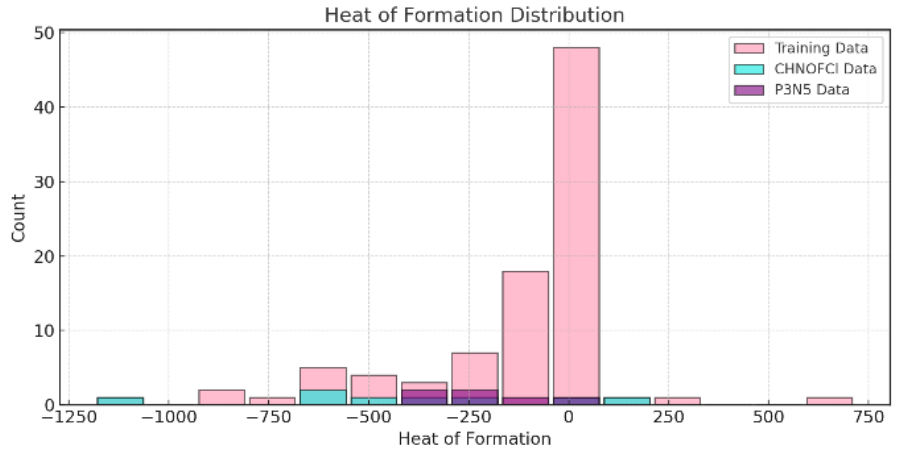
Average Electronegativity
Heat of Formation

RoseAtoms
Model 1

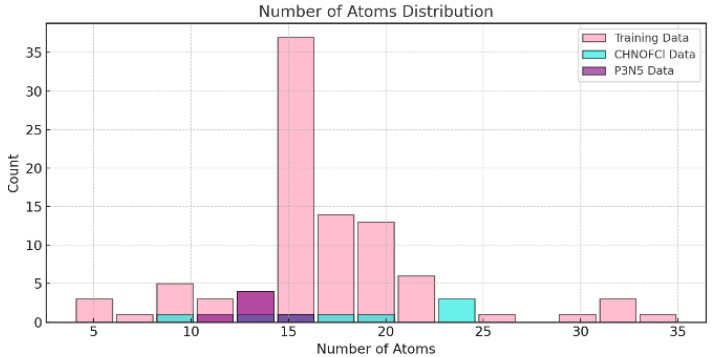
Model Framework: Gaussian Process Regression (GPR) with advanced kernel functions (Matern, RBF, Rational Quadratic, White) to model complex relationships and uncertainties.
Feature Scaling: Input features and target normalized to [0, 1] for improved convergence and performance.
Hyperparameter Tuning: RandomizedSearchCV optimized model parameters via Negative Log-Marginal Likelihood (NLML).
Evaluation: 10-fold cross-validation

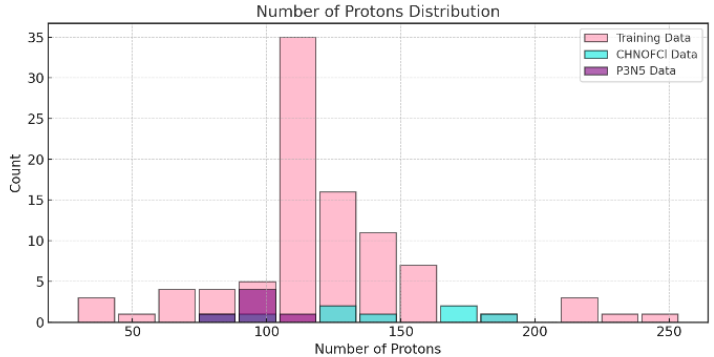
Detonation Velocity

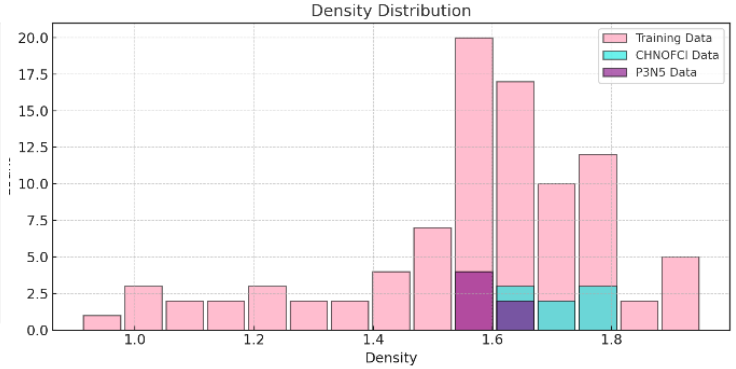


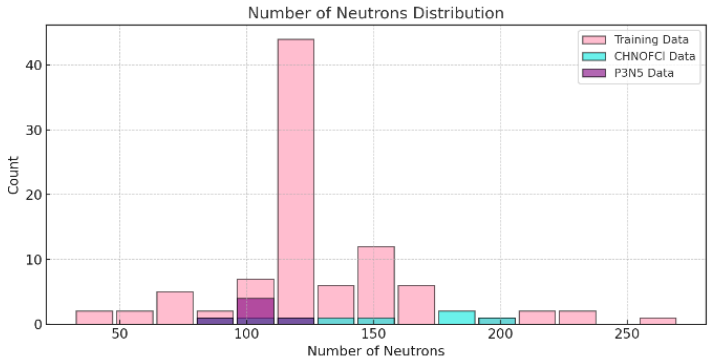


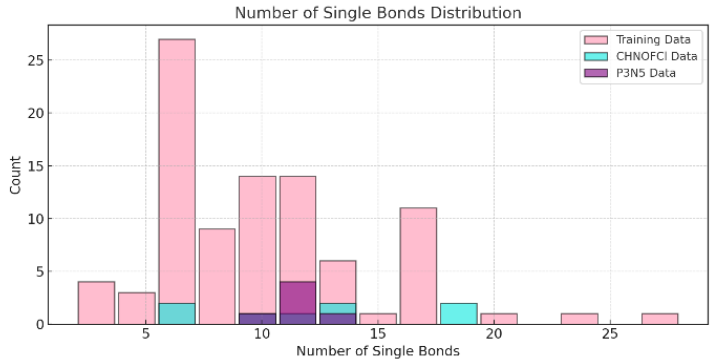
Feature Distribution

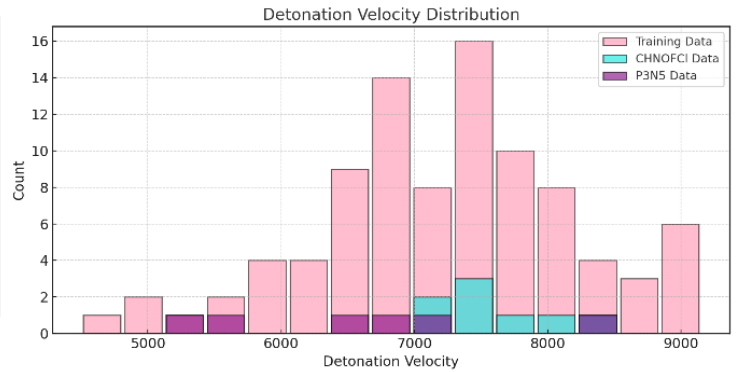


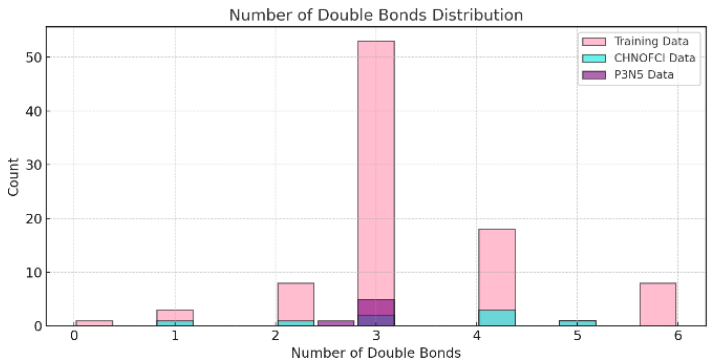


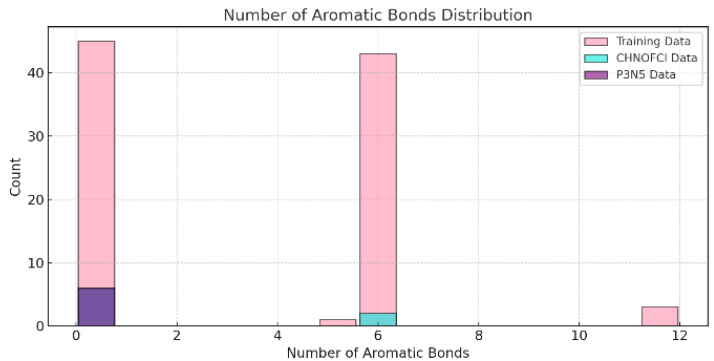










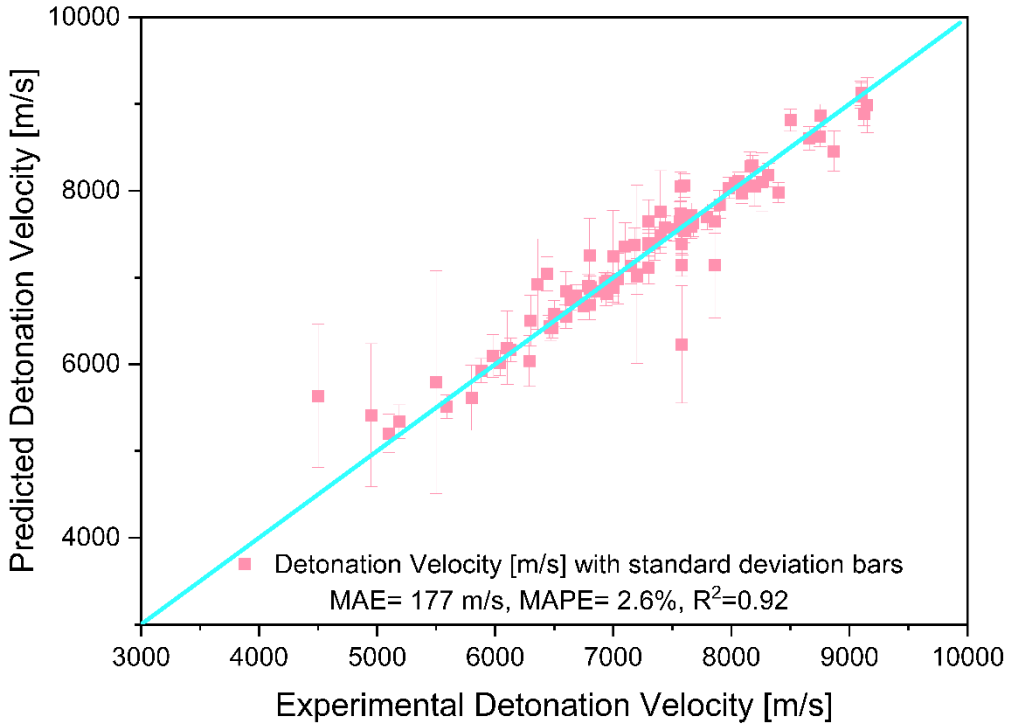


Detonation velocities – removing the handcuffs of the periodic table

Features:
- Average Electronegativity
- Molecular Weight
- Density
- Heat of Formation
- Number of Atoms
- Number of Protons
- Number of Neutrons
- Number of Single Bonds
- Number of Double Bonds
- Number of Aromatic Bonds

+

Gaussian Process Regressor



92 experimental CHNO Detonation Velocities

10 K-Fold Validation

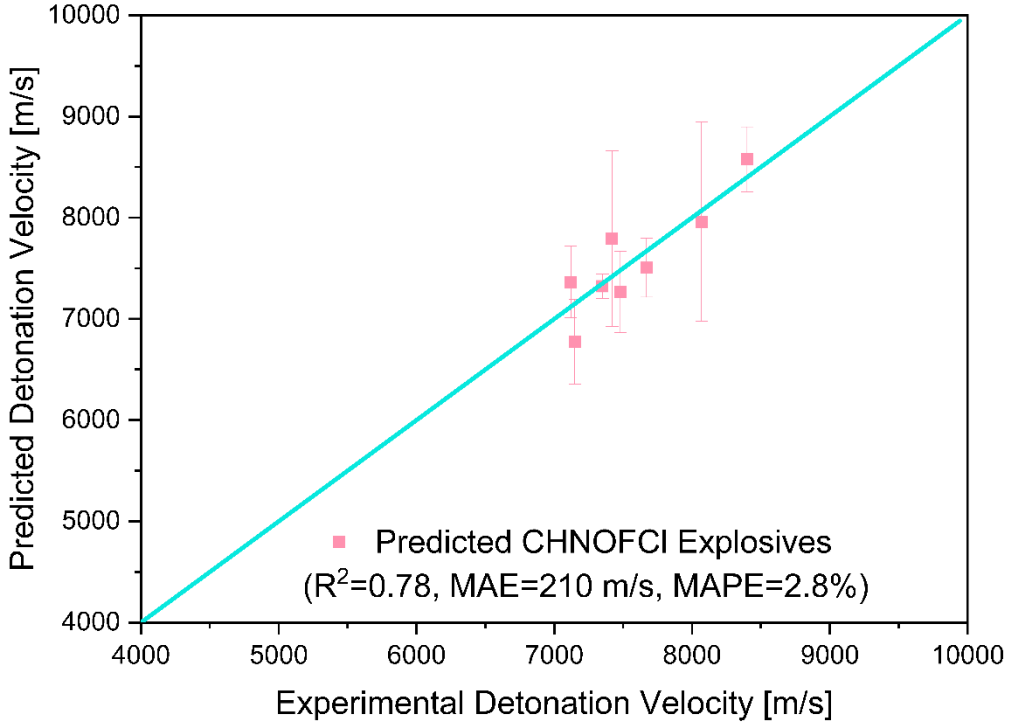
Wahler, S., "Putting an end to the handcuffs of the periodic table: Using atomic values for predicting detonation velocity", Accepted Manuscript, Propellants, Explosives, Pyrotechnics, 2025

Detonation velocities – removing the handcuffs of the periodic table

Features:
- Average Electronegativity
- Molecular Weight
- Density
- Heat of Formation
- Number of Atoms
- Number of Protons
- Number of Neutrons
- Number of Single Bonds
- Number of Double Bonds
- Number of Aromatic Bonds

+

Gaussian Process Regressor



9 experimental CHNOFCl Detonation Velocities

Leave-Out Test Set

Wahler, S., “Putting an end to the handcuffs of the periodic table: Using atomic values for predicting detonation velocity”, Accepted Manuscript, Propellants, Explosives, Pyrotechnics, 2025

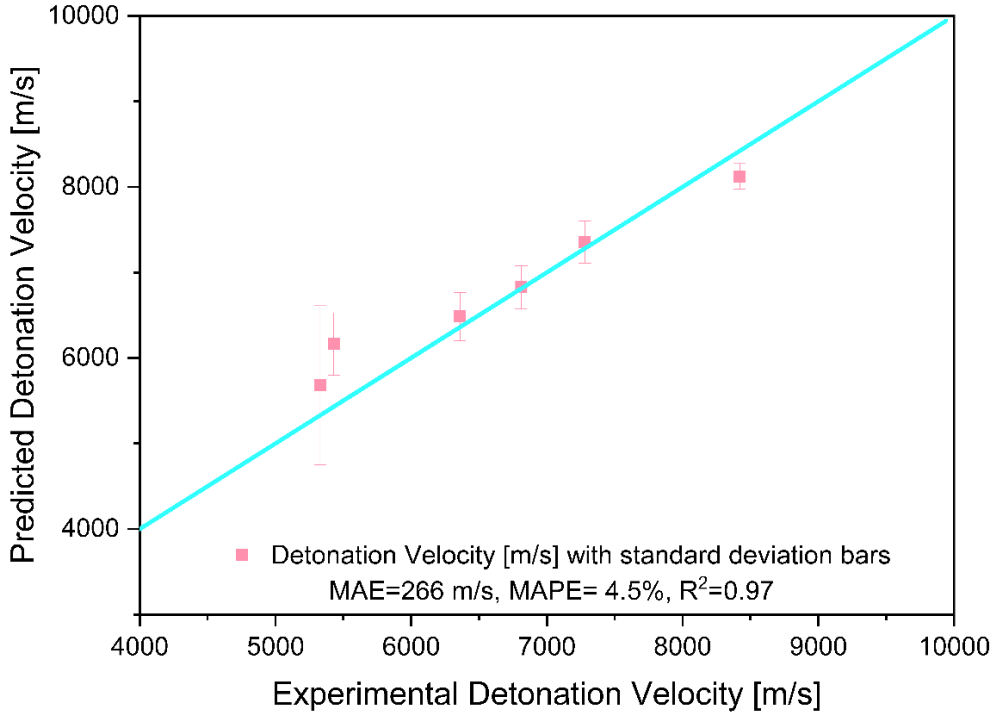
Detonation velocities – removing the handcuffs of the periodic table

Features:

- ~~Average Electronegativity~~
- Molecular Weight
- Density
- ~~Heat of Formation~~
- Number of Atoms
- Number of Protons
- Number of Neutrons
- Number of Single Bonds
- Number of Double Bonds
- Number of Aromatic Bonds

+

Gaussian Process Regressor



Leave-Out Test Set

7 experimental P_3N_5 / Al/ CHNO Mixtures Detonation Velocities

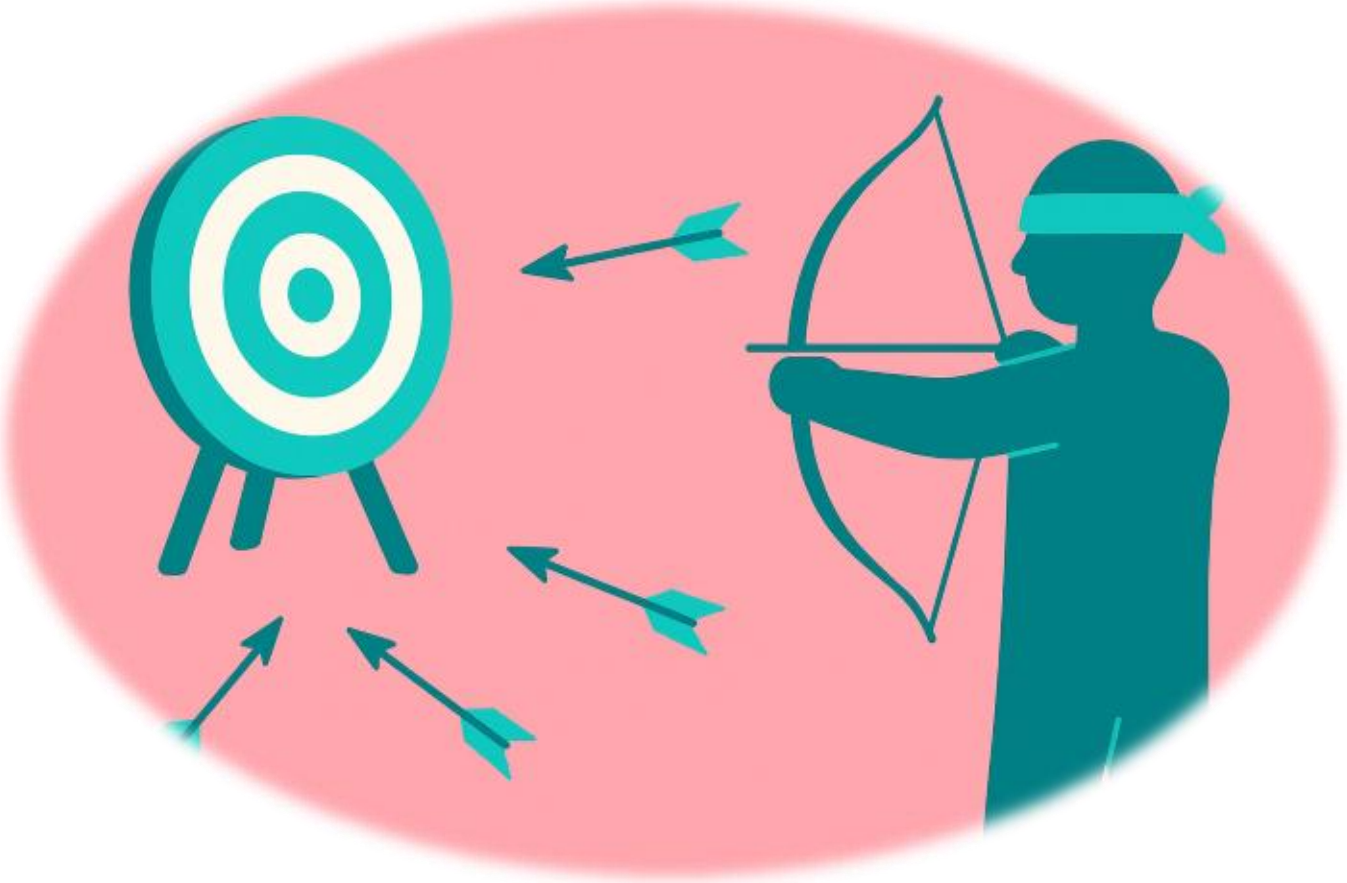
Wahler, S., "Putting an end to the handcuffs of the periodic table: Using atomic values for predicting detonation velocity", Accepted Manuscript, Propellants, Explosives, Pyrotechnics, 2025

What is Gaussian Process Regression? *(For dummies)*

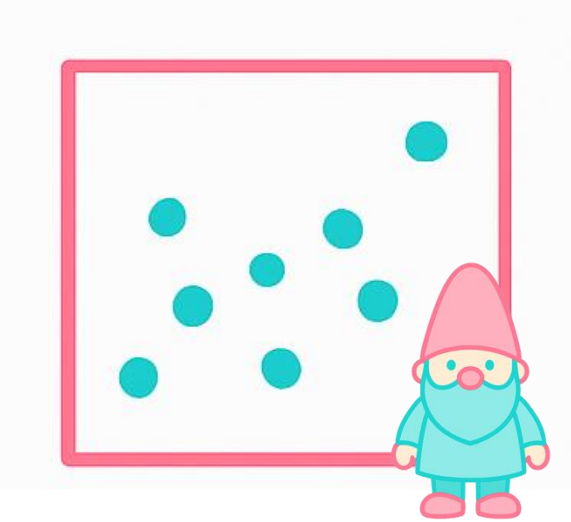
Gaussian Process Regression is like aiming at targets blindfolded…

Confident near where you've already hit — but uncertain where you've never shot.

GPR models this by showing not just a prediction, but **how confident** the model is in different regions.



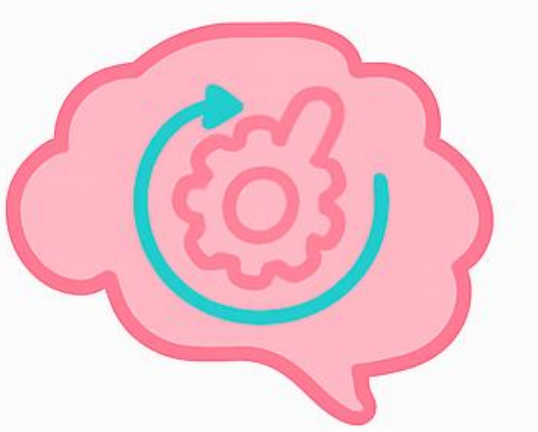
What are Gaussian Process Regressors good for?



Small Datasets



Smooth functions



Bayesian Frame Works



Uncertainty predictions

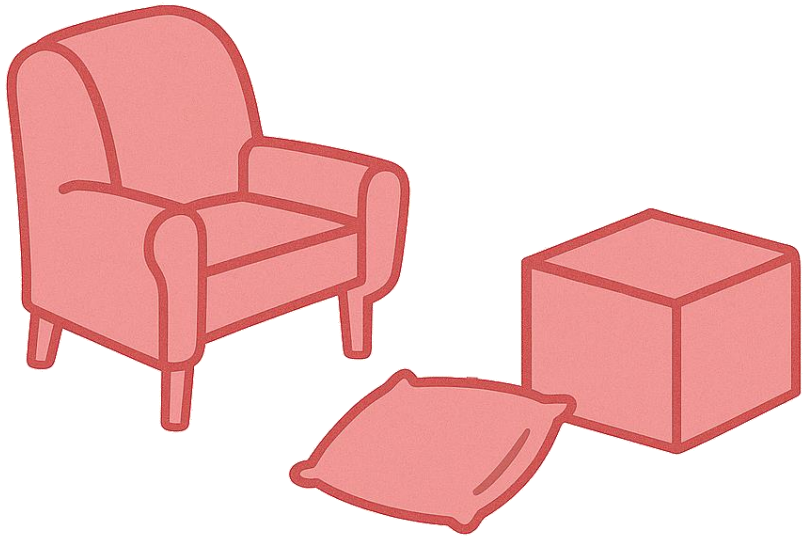
How do we fit the curves?

- Like a blanket over pillows and chairs of different heights
- Less stable/ higher uncertainty in the parts not close to data points

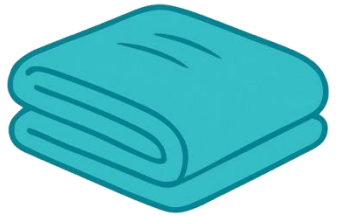


The Data

•Gaussian Process Regression (GPR) scales poorly with **sample size** n, not feature count d
•Works best with **< 1,000 samples**
•Becomes slow with **1,000–5,000 samples**
• Infeasible beyond **10,000**
•Handles moderate feature counts well (up to **~100 features**)

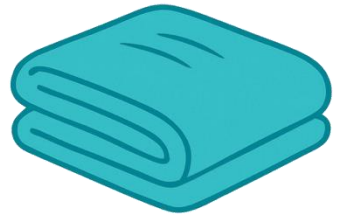


The Function



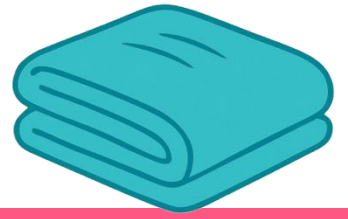
Blanket Part	Mathematical Term	Equations
Shape of blanket	Mean prediction μ(x)	$\mu(x) = k(x, X)^T K(X, X)^{-1} y$
Sag or looseness	Uncertainty σ(x)	$\sigma(x) = \sqrt{k(x, x) - k(x, X)^T K(X, X)^{-1} k(x, X)}$
Blanket Material (stretchiness)	Kernel k(x,x')	$k(x, X) = \exp\left(-\frac{(x - X)^2}{2\ell^2}\right)$
Tightness with chair	Conditioning with K(X,X)	$K(X, X) = [k(x_i, x_j)]_{i,j=1}^{n}$

Kernel in GPR's



Kernel Type	Material Analogy	Description
RBF (SE) Kernel	Stretchy fleece	Very smooth, infinite differentiability
Matern Kernel v = 3/2	thick cotton blanket	soft but with more folds and texture - more realistic for rough data
Matern Kernel v = 5/2	finer cotton or wool	smoother drape, fewer wrinkles
Linear Kernel	Stiff tarp	No curve — straight-line dependencies
Periodic Kernel	Quilted fabric pattern	Repeats patterns like waves or stitches

Kernel in GPR's

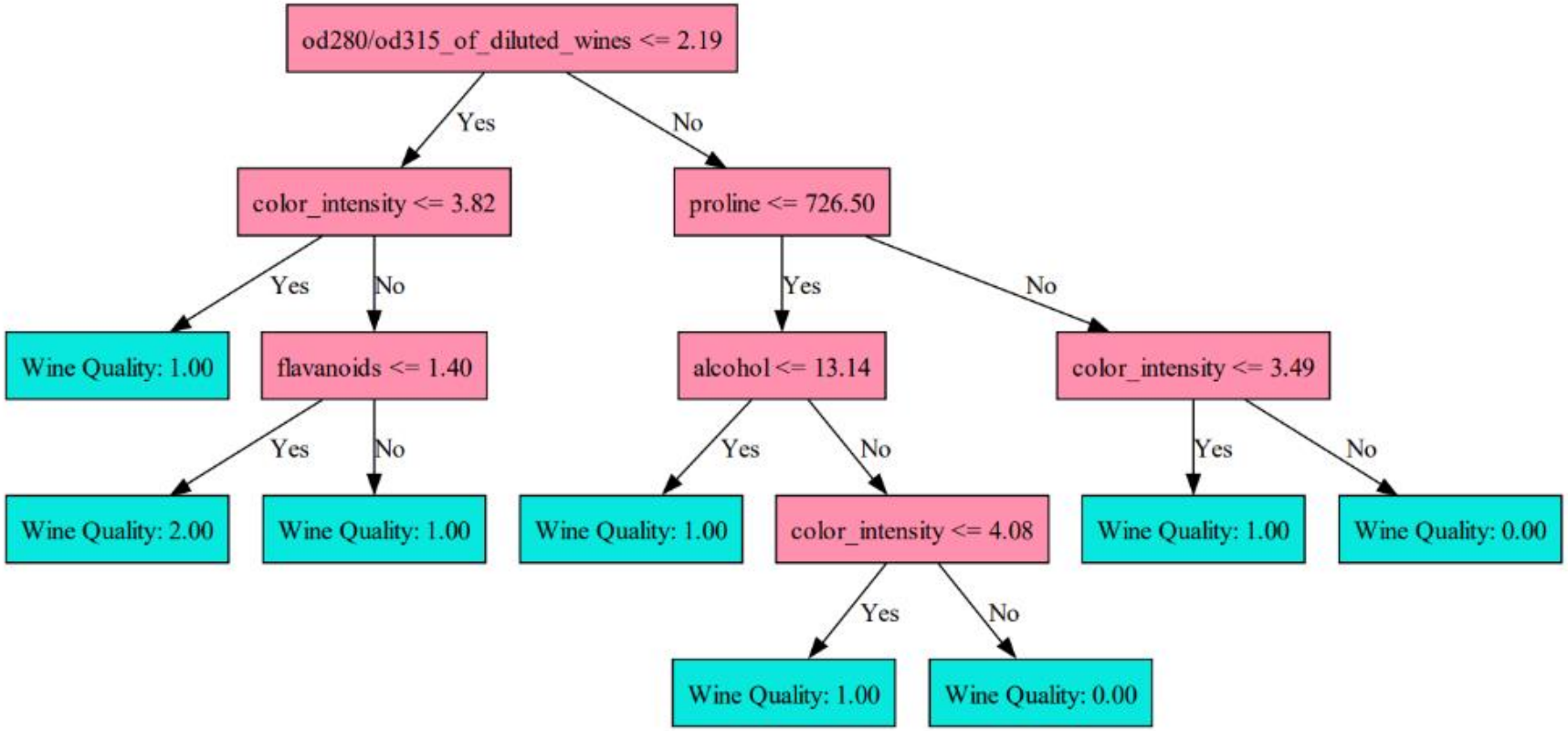


Kernel Type	Description	
RBF (SE) Kernel	Very smooth, infinite differentiability	$k(x,x') = \sigma_f^2 \exp\left(-\frac{(x-x')^2}{2\ell^2}\right)$
Matern Kernel v = 3/2	Less smooth, more realistic for rough data	$k(x,x') = \sigma_f^2 \left(1+\frac{\sqrt{3}\lvert x-x'\rvert}{\ell}\right)\exp\left(-\frac{\sqrt{3}\lvert x-x'\rvert}{\ell}\right)$
Matern Kernel v = 5/2	Less smooth, more realistic for rough data	$k(x,x') = \sigma_f^2 \left(1+\frac{\sqrt{5}\lvert x-x'\rvert}{\ell}+\frac{5(x-x')^2}{3\ell^2}\right)\exp\left(-\frac{\sqrt{5}\lvert x-x'\rvert}{\ell}\right)$
Linear Kernel	No curve — straight-line dependencies	$k(x,x') = \sigma_b^2 + \sigma_v^2(x-c)(x'-c)$
Periodic Kernel	Repeats patterns like waves or stitches	$k(x,x') = \sigma_f^2 \exp\left(-\frac{2\sin^2(\pi\lvert x-x'\rvert/p)}{\ell^2}\right)$

Tree Based Models



Decision Tree Models

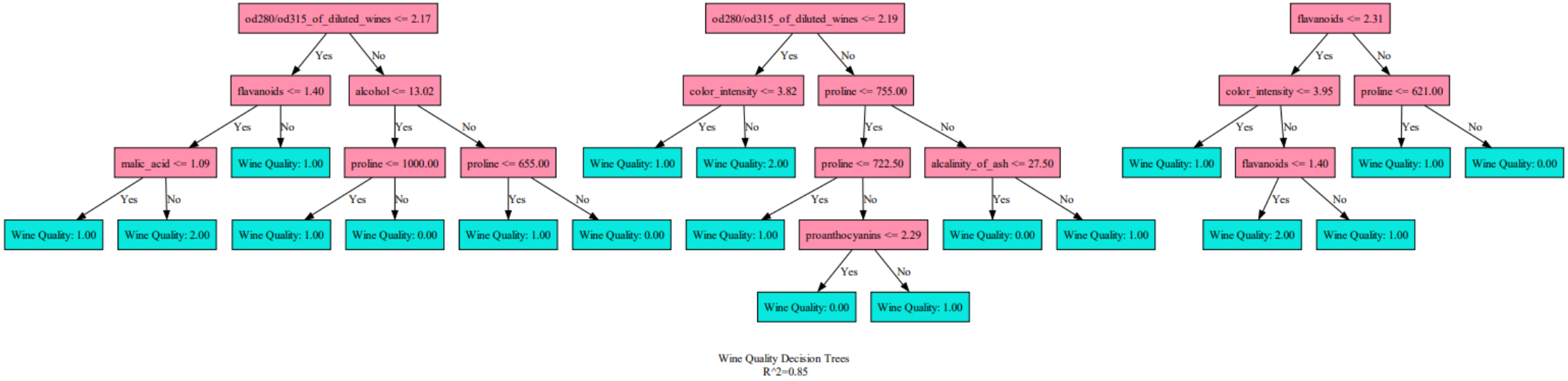


Wine Quality Decision Tree
R^2=0.71

How do I “grow” my tree?

Hyperparameter	Effect
max_depth	Limits tree depth → controls model complexity
min_samples_split	Minimum samples to split a node → avoids overfitting
min_samples_leaf	Minimum samples at leaf → helps generalization
max_features	Limits features to consider at each split → adds randomness
criterion	Split metric (e.g., Gini, entropy for classification)

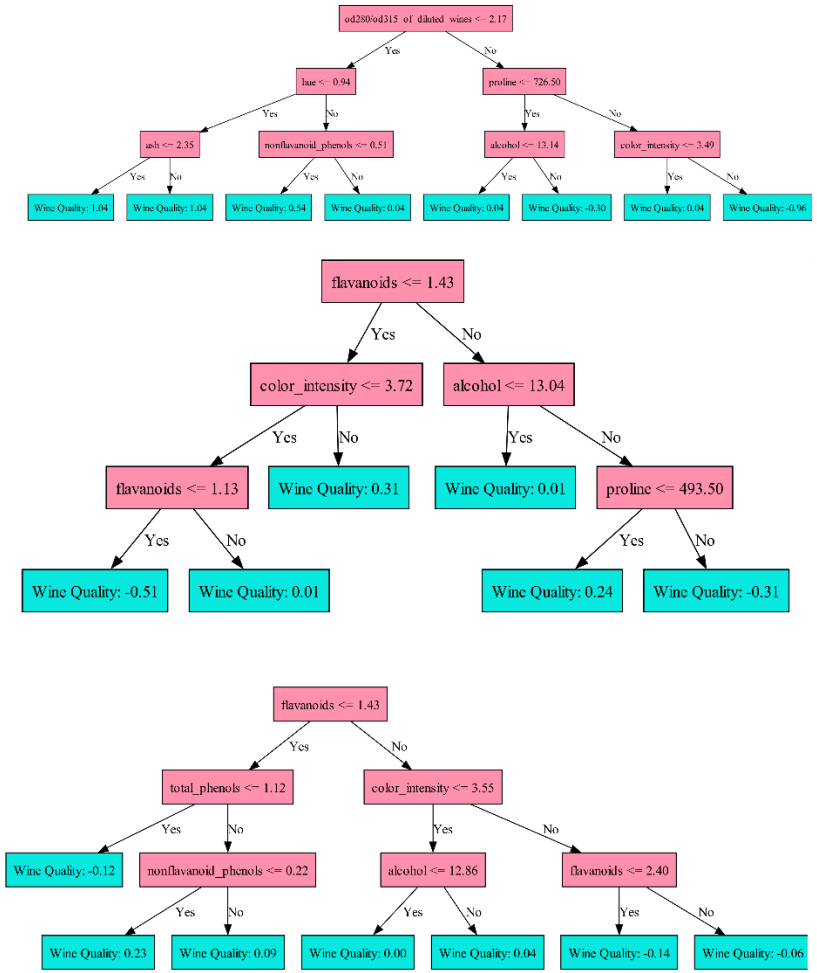
Random Forrest Model



Growing a forest

Hyperparameter	Effect
n_estimators	Number of trees: more trees = better but slower
bootstrap	Whether to sample data with replacement
max_samples	Subsample size per tree (if bootstrap is true)
max_features	How many features to try per split: helps diversity

Gradient Boosted Regressor

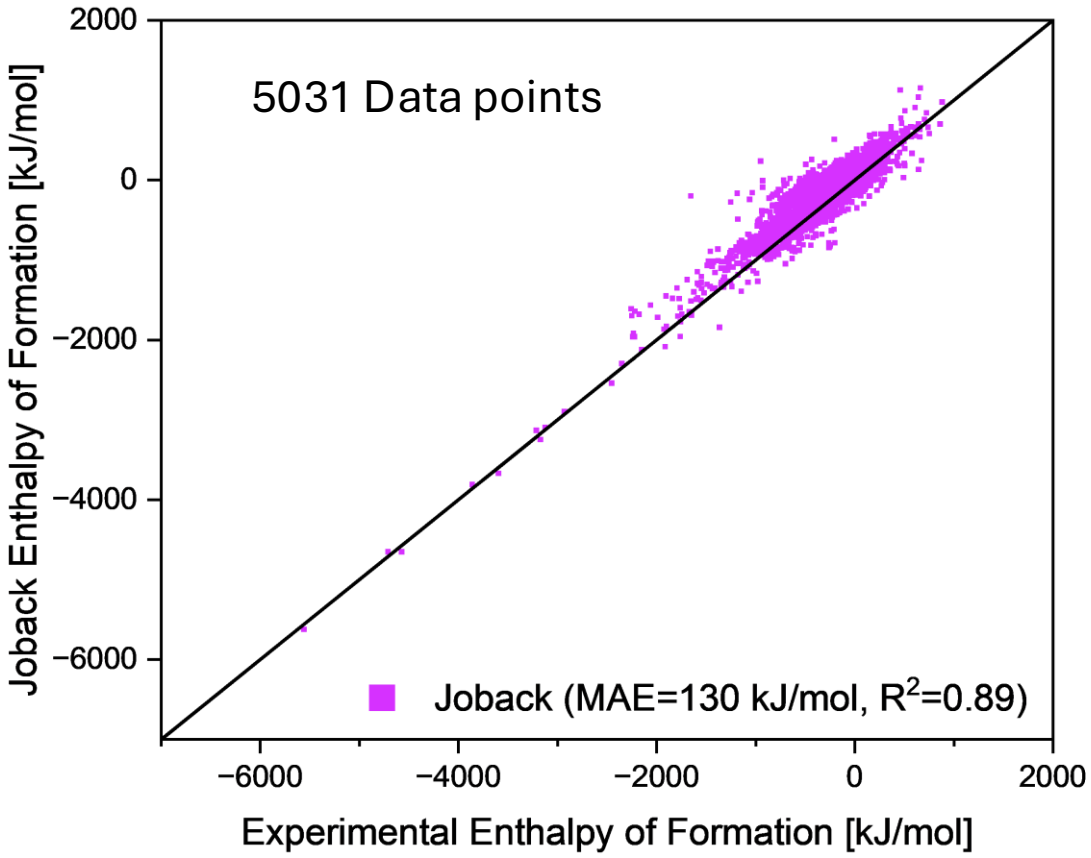


Wine Quality Decision Trees
R^2=0.82

How to boost

Hyperparameter	Effect
n_estimators	Number of boosting rounds
learning_rate	Shrinks contribution of each tree → lower = better, but slower
max_depth	Depth of each tree
subsample	Fraction of data to use per tree → combats overfitting
colsample_bytree	Fraction of features used per tree
gamma / min_split_loss	Minimum loss reduction to make a split
reg_alpha, reg_lambda	L1 / L2 regularization on leaf weights

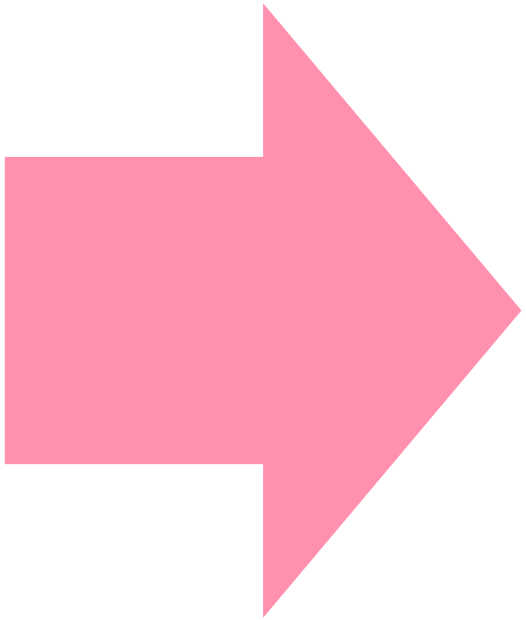
Parametization of empirical approximations using statistical modelling (PEAS)



Wahler, S.*, Reparametrizing Joback's group additivity theory for thermochemical predictions – Thermochemical building blocks for targeted design of new compounds, 2024, Propellants, Explosives and Pyrotechnics, 1-15



130 kJ





40 kJ

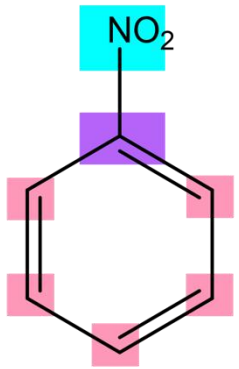
Wahler, S.*, Reparametrizing Joback's group additivity theory for thermochemical predictions – Thermochemical building blocks for targeted design of new compounds, 2024, Propellants, Explosives and Pyrotechnics, 1-15

Empirical Methods

Machine Learning

EoF,
Density, …

5 x C-H (aromatic)
1 x C (aromatic)
1 x –NO_2



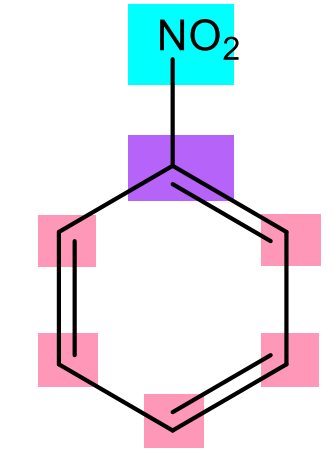
- Scaling of the data
- Hyperparameter Tuning
- Model

EoF,
Density, …

Wahler, S.*, Reparametrizing Joback's group additivity theory for thermochemical predictions – Thermochemical building blocks for targeted design of new compounds, 2024, Propellants, Explosives and Pyrotechnics, 1-15

Assembly of the Models

Featurization



Joback

Naef

Benson

Reshavatz

To specific, limiting to fewer compounds

Scaling

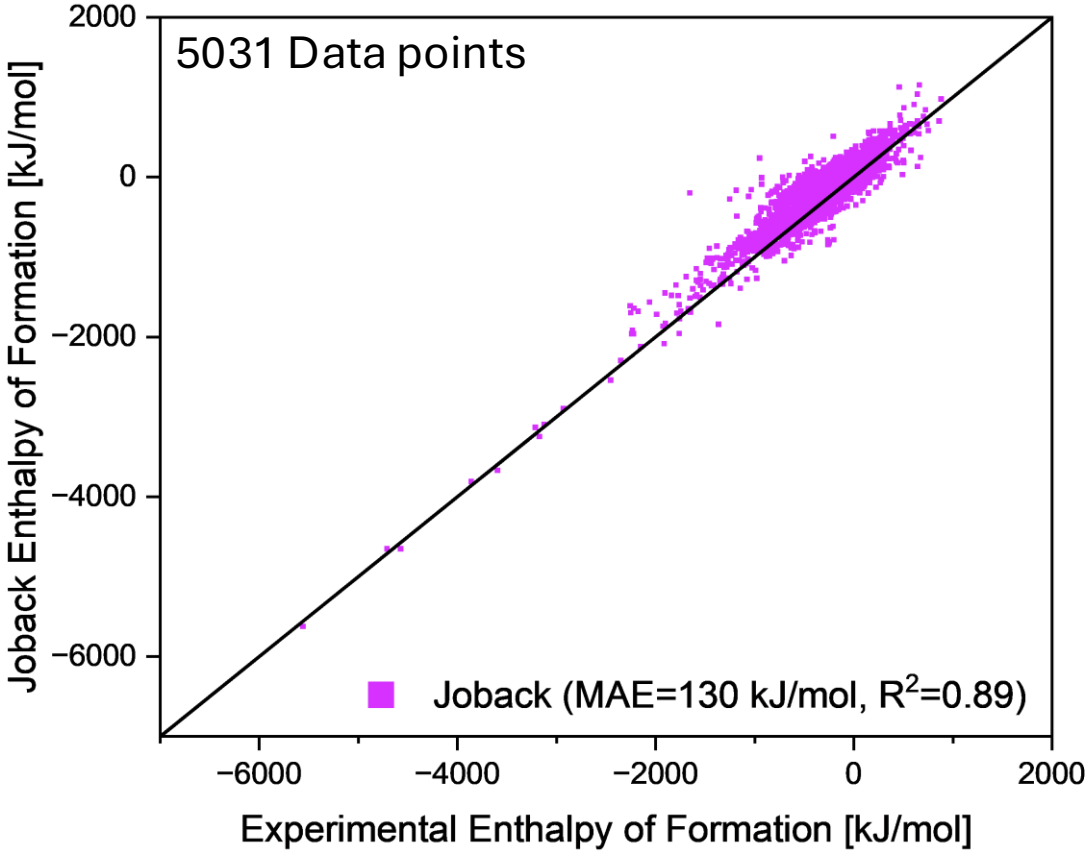
MinMaxScaler

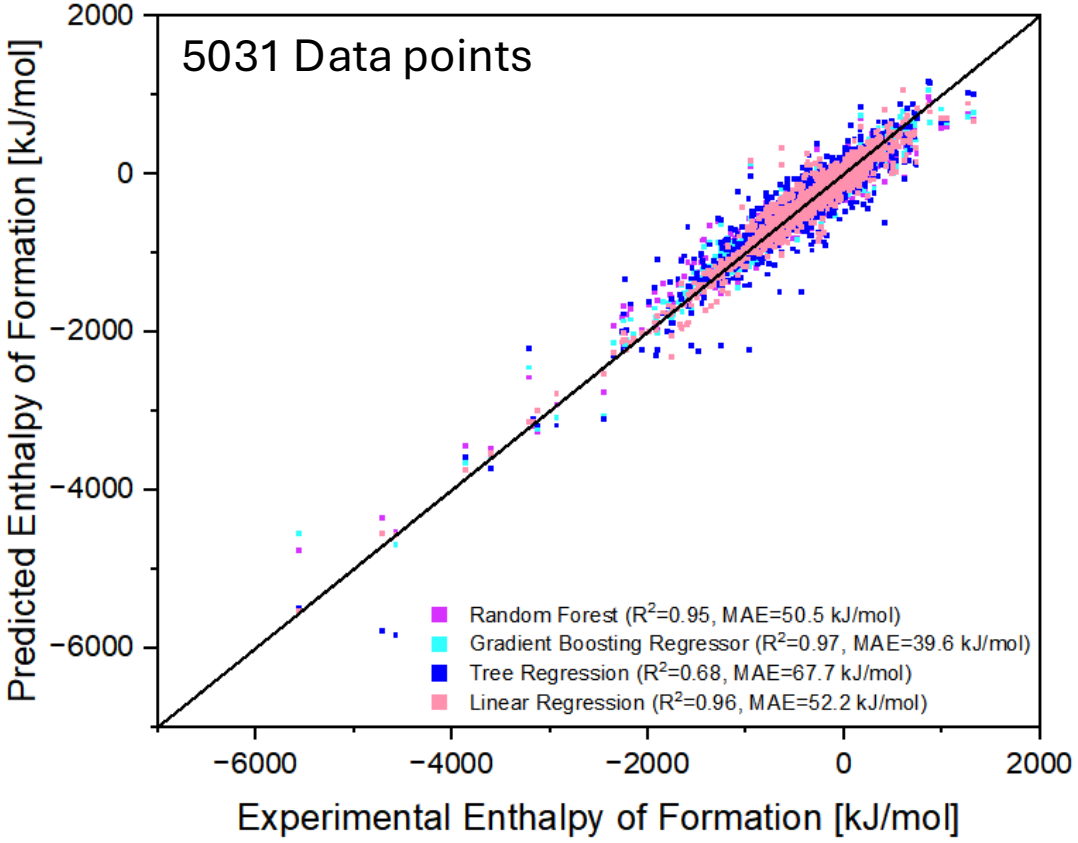
Model and Hyperparameter tuning

Model	Hyperparameters tuned with GridSearchCV
Random Forest Model	'n_estimators': [50, 100, 200, 500], 'max_depth': [5, 10, 20, None], 'min_samples_split': [2, 5, 10], 'min_samples_leaf': [1, 2, 4]]
Gradient Boosting Regressor	'learning_rate': [0.01, 0.1, 1], 'n_estimators': [50, 100, 200, 500], 'max_depth': [5, 10, 20, None], 'min_samples_split': [2, 5, 10], 'min_samples_leaf': [1, 2, 4], 'max_features': ['auto', 'sqrt', 'log2']
Decision Tree Regressor	'dt__criterion': ['mse', 'mae'], 'dt__max_depth': [3, 5, 7, None], 'dt__min_samples_split': [2, 5, 10], 'dt__min_samples_leaf': [1, 2, 4], 'dt__max_features': ['sqrt', 'log2', None]
Linear Regression	'fit_intercept': [True, False], 'normalize': [True, False]

Wahler, S.*, Reparametrizing Joback's group additivity theory for thermochemical predictions – Thermochemical building blocks for targeted design of new compounds, 2024, Propellants, Explosives and Pyrotechnics, 1-15

Parametization of empirical approximations using statistical modelling (PEAS)

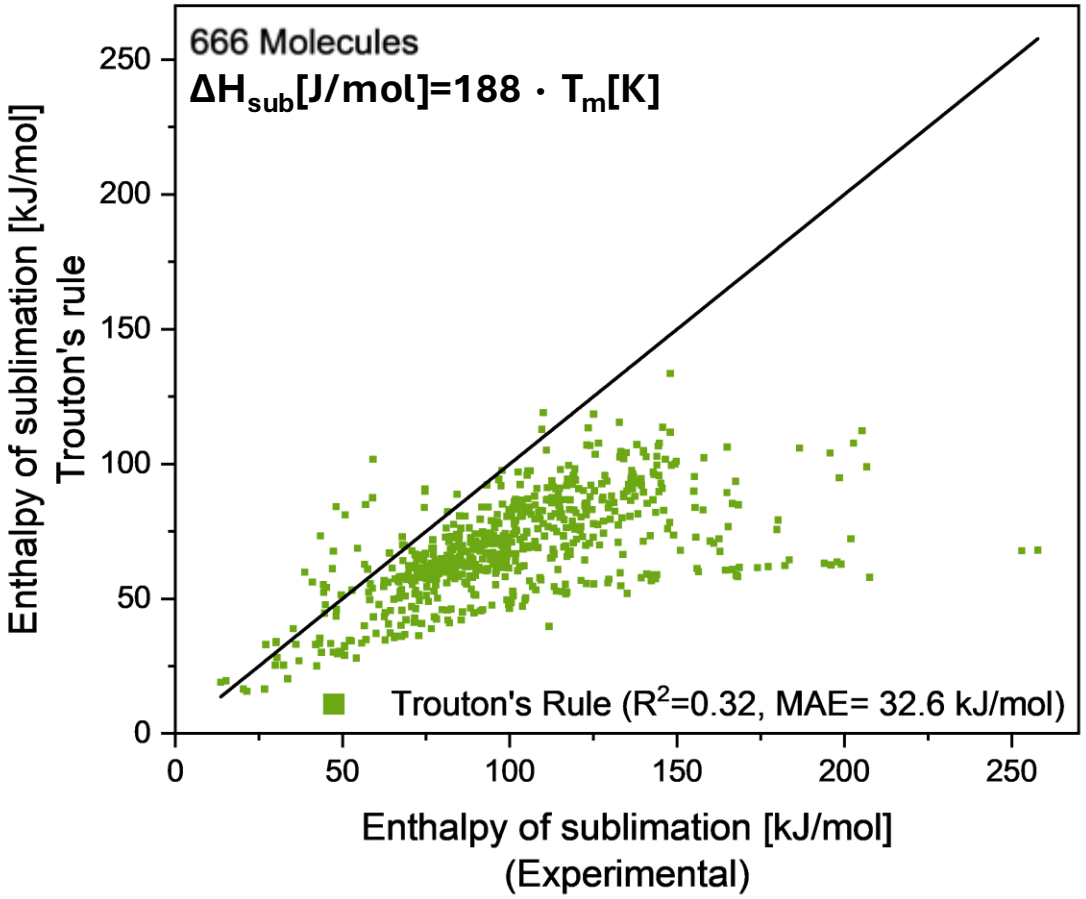


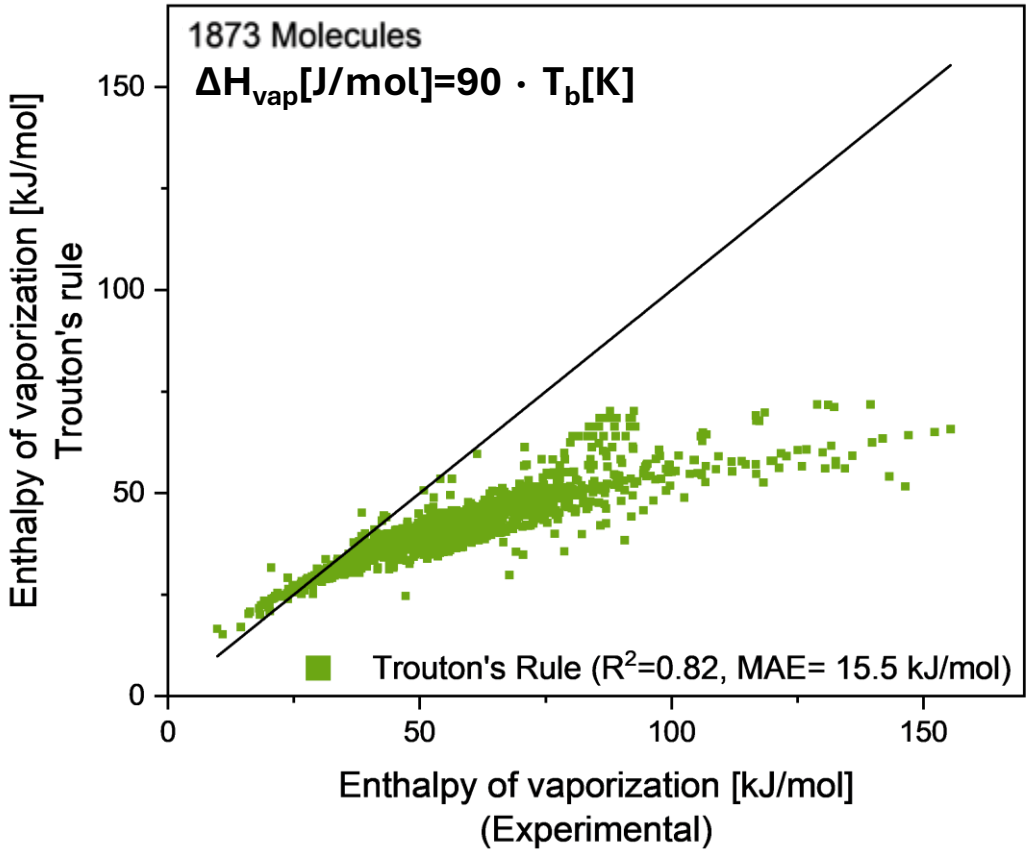


Showing the test sets of 10 K-Fold Validation!

Wahler, S.*, Reparametrizing Joback's group additivity theory for thermochemical predictions – Thermochemical building blocks for targeted design of new compounds, 2024, Propellants, Explosives and Pyrotechnics, 1-15

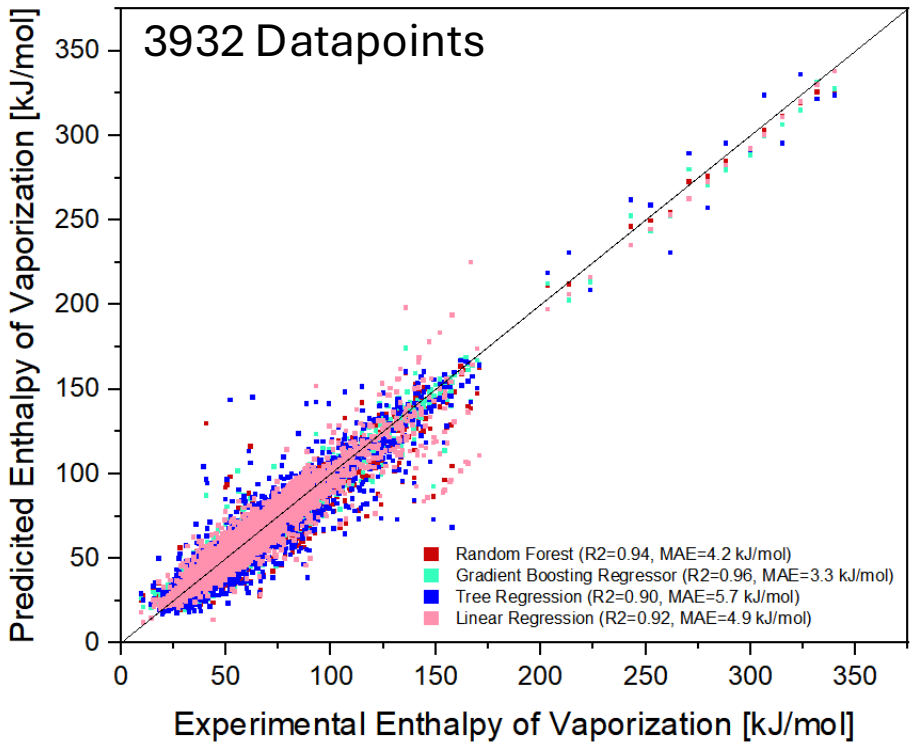
Parametization of empirical approximations using statistical modelling (PEAS)

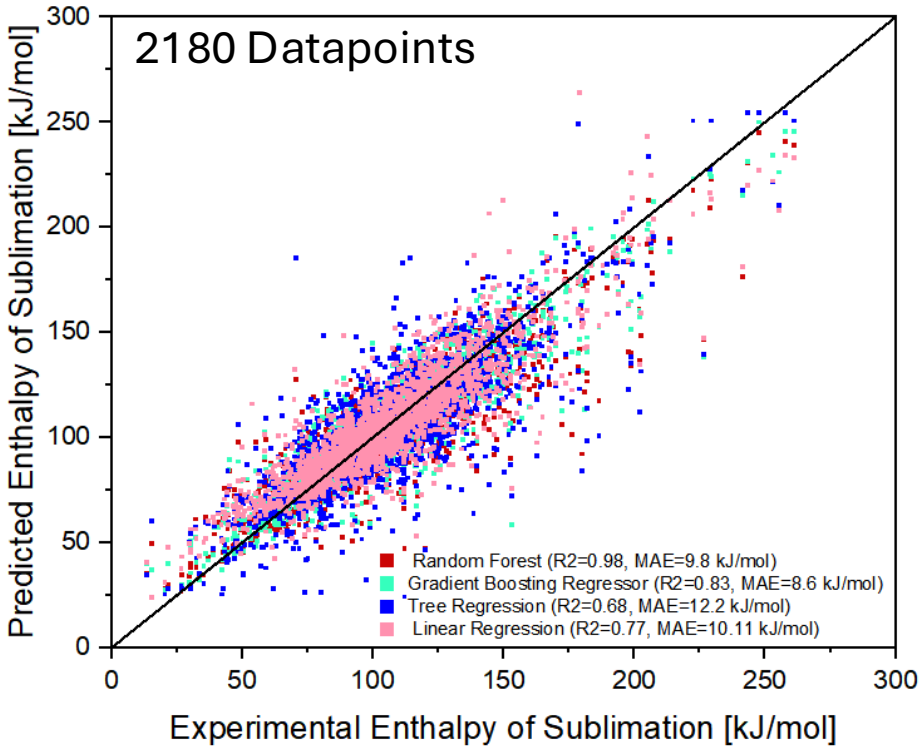




Westwell, M. S.; Searle, M. S.; Wales, D. J.; Williams, D. H., J. Am. Chem. Soc. 1995, 117, 5013.
Wahler, S., Klapötke, T.M., Chung., P, *J. Energ. Mater.* **2023**, Accepted Manuscript

Parametization of empirical approximations using statistical modelling (PEAS)





Showing the test sets of 10 K-Fold Validation!

Wahler, S.*, Reparametrizing Joback's group additivity theory for thermochemical predictions – Thermochemical building blocks for targeted design of new compounds, 2024, Propellants, Explosives and Pyrotechnics, 1-15

Linear Regression, GPR's and Decision Tree's



- More Complicated Model = Not always better
- Can assist in efficiently estimating a property efficiently
- GPR: Remember chairs and blankets
- Decision trees: Random and Subsequent

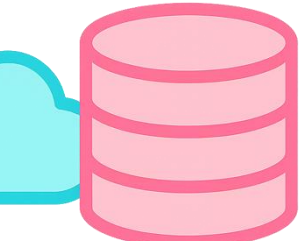
Key Takeaways

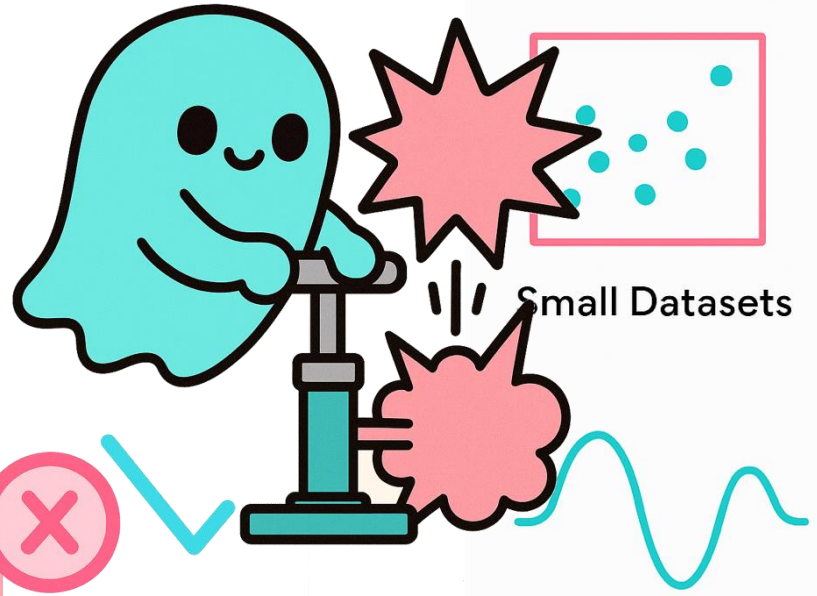
- Supervised learning = learning from labeled chemistry data; how you evaluate it matters as much as the model itself
- Split by scaffold, not randomly — interpolation is easy, extrapolation isn't
- Prediction error = bias + variance + noise; learning curves show you which one is hurting you
- Parameters are learned; hyperparameters are tuned on validation data, never on test data
- Data bias becomes model bias — know how your dataset was built

Using AI to Talk about AI









Small Datasets



Uncertainty Matters

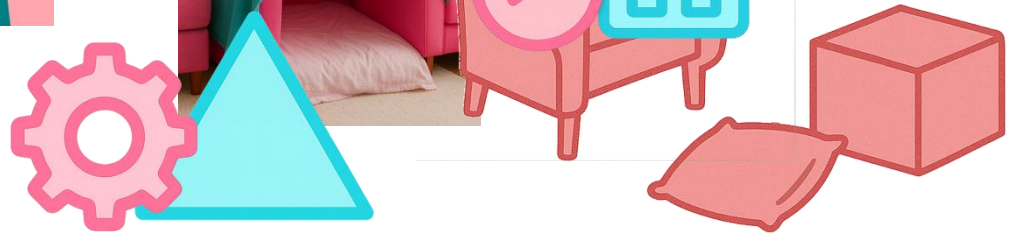


Bayesian Framewock

Good thing I trust these “AI Framewocks” to correct my spelling and grammar as non-native speaker



Smooth, Continuous Functions



Questions?

Dr. Sabrina Wahler

swahler@caltech.edu