



The views expressed in this presentation are those of the authors and do not necessarily reflect the views or policies of the U.S. EPA

Comparative Analysis of Applicability Domain Methods

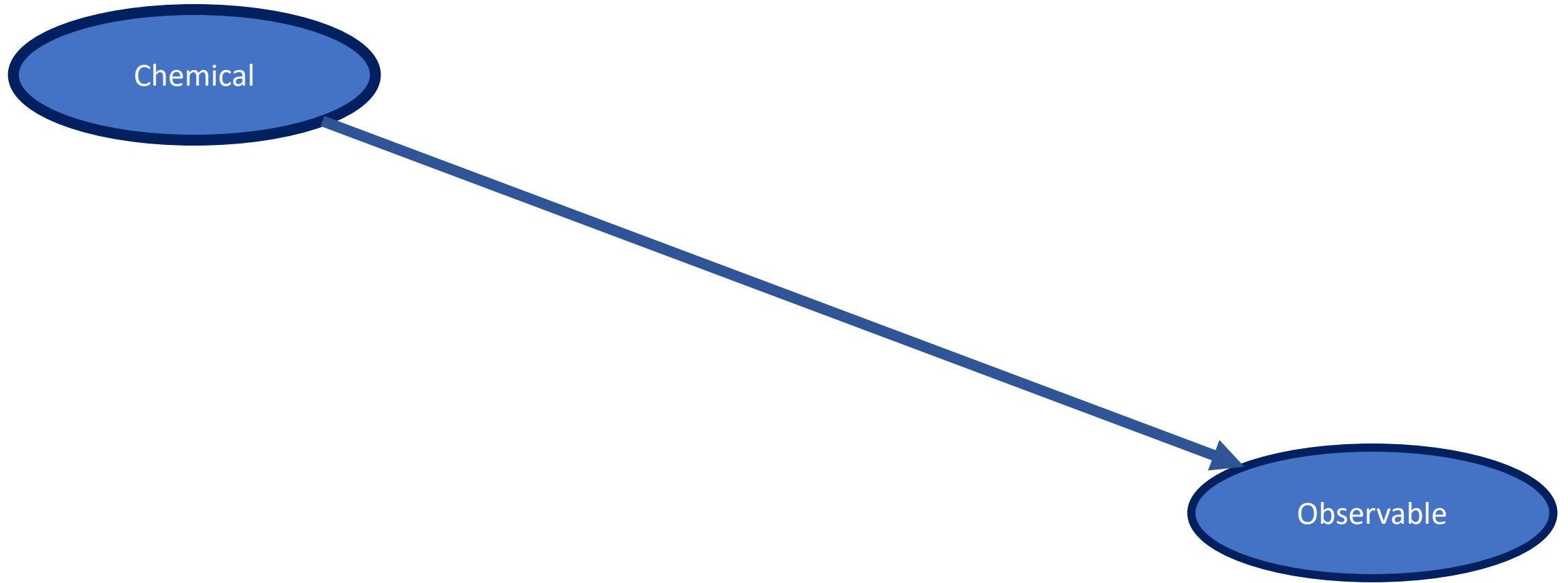
Nathaniel Charest, Christian Ramsland, Todd Martin, Antony Williams



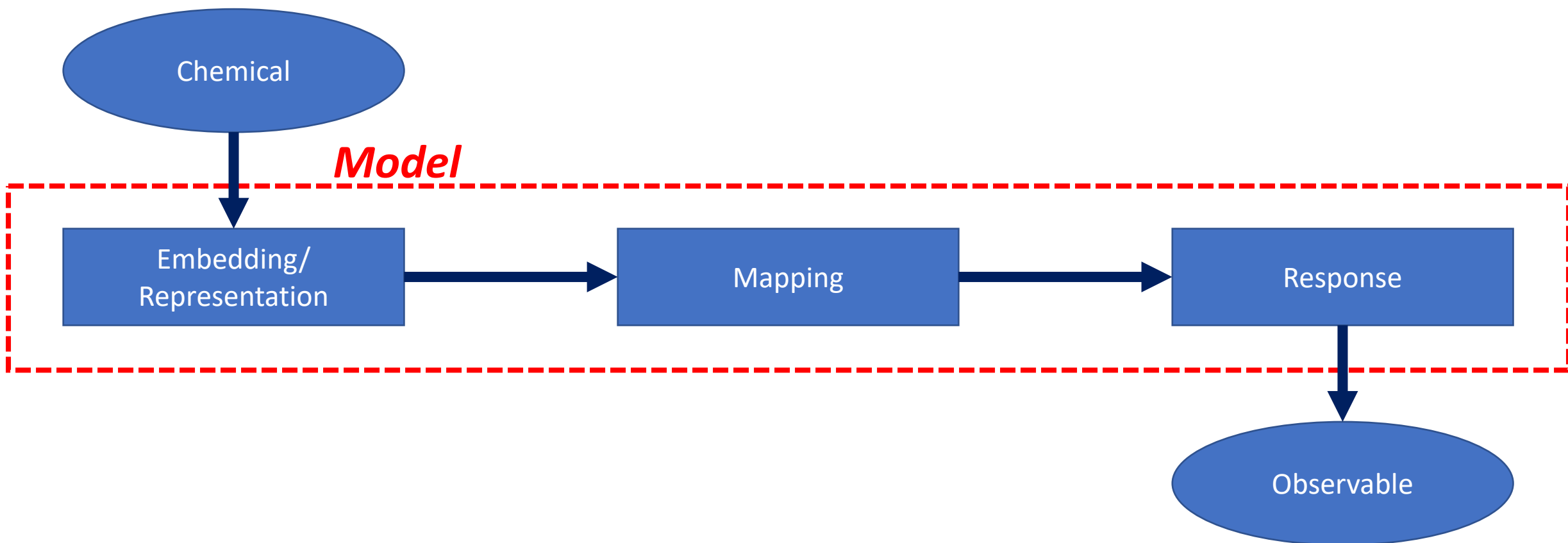
OECD Principle 3: Definition of an Applicability Domain

- Inputs for which a structure-activity model is presumed to be more accurate, or “applicable”
- Can be qualitative
 - “Any organic molecule that weighs less than 1000 amu”
- Quantitative applicability domains are preferred, as they can be programmatically integrated

What is an Applicability Domain?

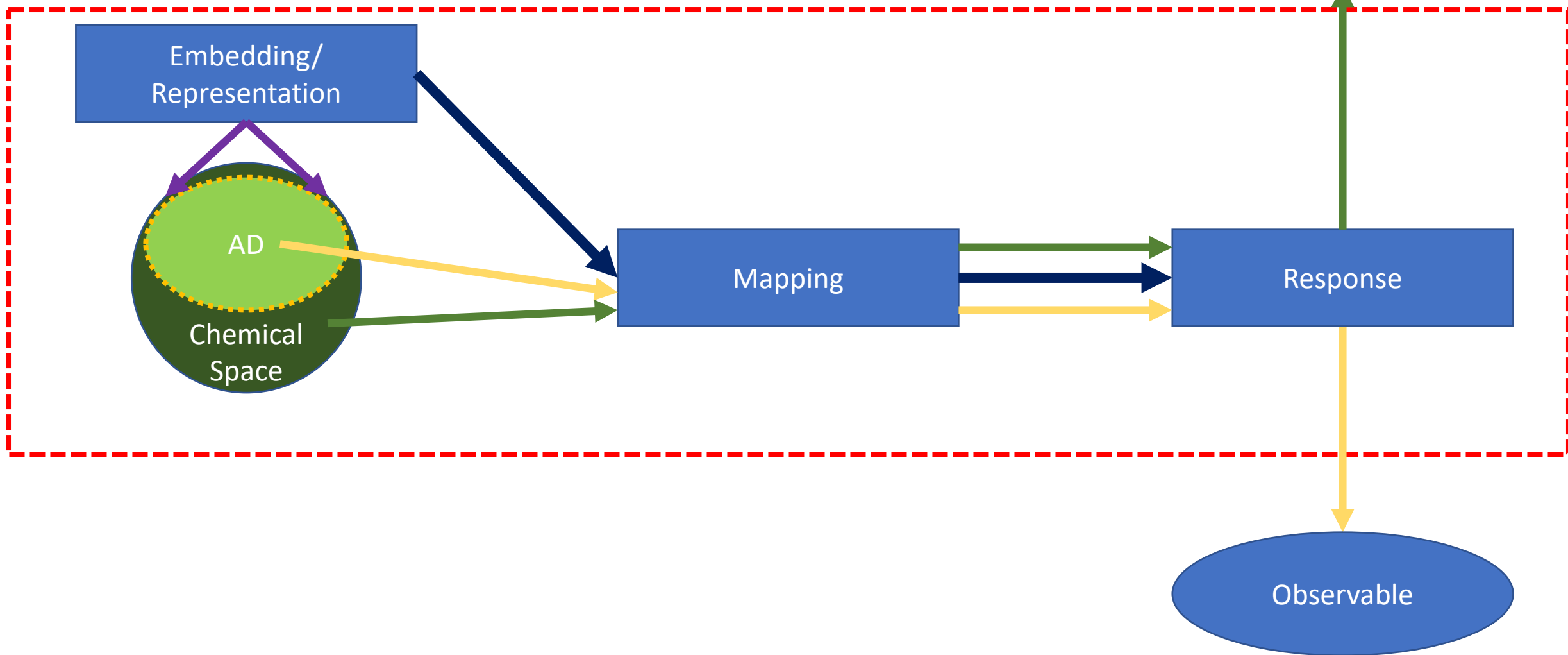


What is an Applicability Domain?



What is an Applicability Domain?

Model



What is an Applicability Domain (AD)?

- A partitioning of input (chemical) space that determines whether the associated mapping agrees with reality or produces an erroneous prediction
- What measures of chemical similarity will be most relevant to assessing model applicability outside the training set?
- We need methods to assess a similarity, or applicability, measure and compare them to each other with objective performance metrics

Evaluating An Applicability Measure

- Indices of Interest
 - OPERA Local Index
 - Weighted Euclidean Similarity
 - Average Standardized Cosine Similarity of 3-Nearest Neighbors
- $\frac{I}{O} Ratio = \frac{R_{accepted}^2}{R_{rejected}^2}$
- Endpoints
 - OPERA Dataset
 - Vapor Pressure
 - LogKmHL – Biotransformation half-life in zebra fish
 - TEST Datasets
 - IGC50 – *Tetrahya pyriformis* 50% growth inhibition concentration
 - LD50 – Rat Lethal Dose 50%

EPA Models

Mansouri et al. *J Cheminform* (2018) 10:10
<https://doi.org/10.1186/s13321-018-0263-1>

 Journal of Cheminformatics

RESEARCH ARTICLE

Open Access



OPERA models for predicting physicochemical properties and environmental fate endpoints

Kamel Mansouri^{1,2,3*} , Chris M. Grulke¹, Richard S. Judson¹ and Antony J. Williams¹

EPA Models

TEST (Toxicity Estimation Software Tool) Ver 4.1

Citation:

Martin, T., P. Harten, AND D. Young. TEST (Toxicity Estimation Software Tool) Ver 4.1. U.S. Environmental Protection Agency, Washington, DC, EPA/600/C-12/006, 2012.

Impact/Purpose:

The purpose of the TEST software is to estimate toxicity values and physical properties for chemicals from their molecular structure. The estimated toxicity and physical property values can be used to provide data needed for improving the sustainability of chemical process designs (using algorithms such as the WASTE Reduction Algorithm) or in the selection of greener solvents (for example the Paris III software).

Contact

[SHERRY PARKER](#)

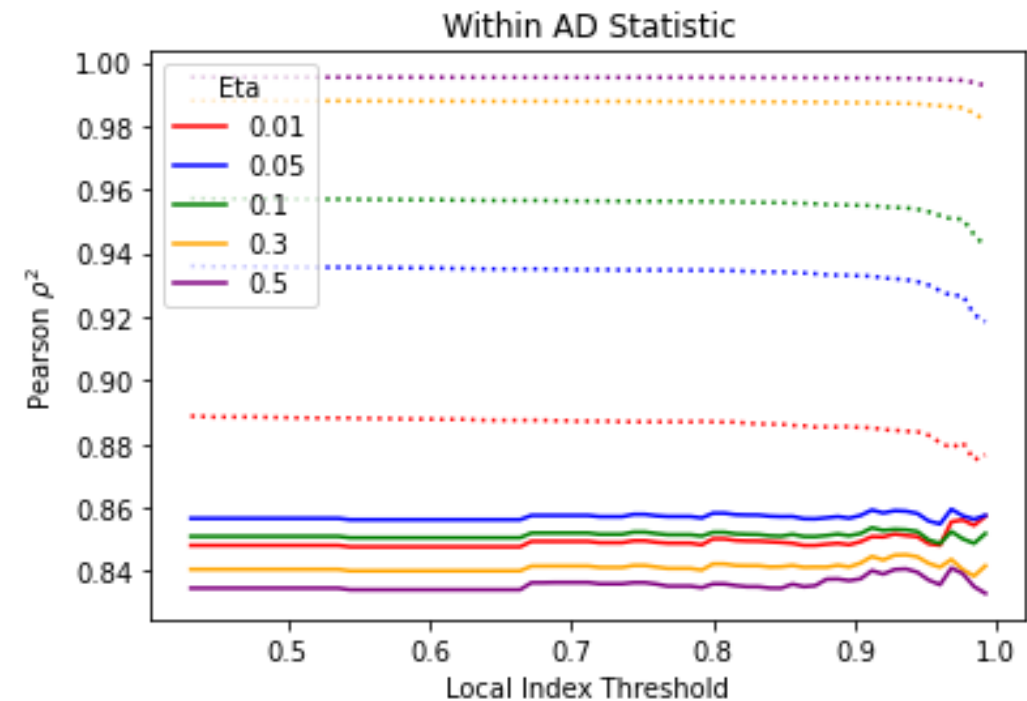
phone: 513-487-2830

fax: 5135697471

email: parker.sherry@epa.gov

Evaluating An Applicability Measure

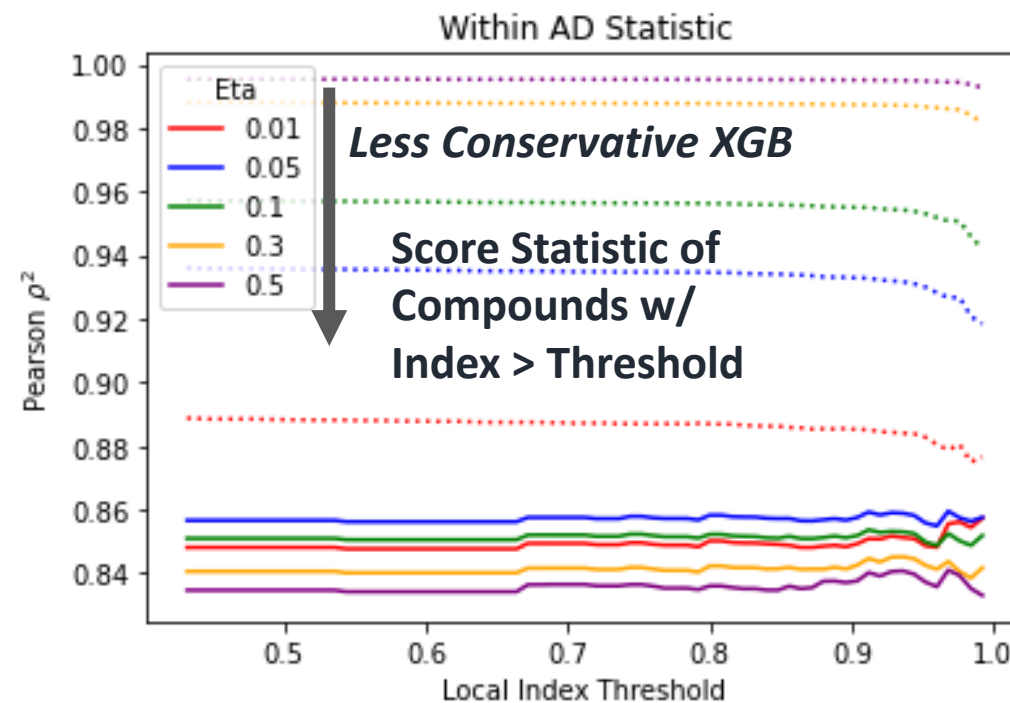
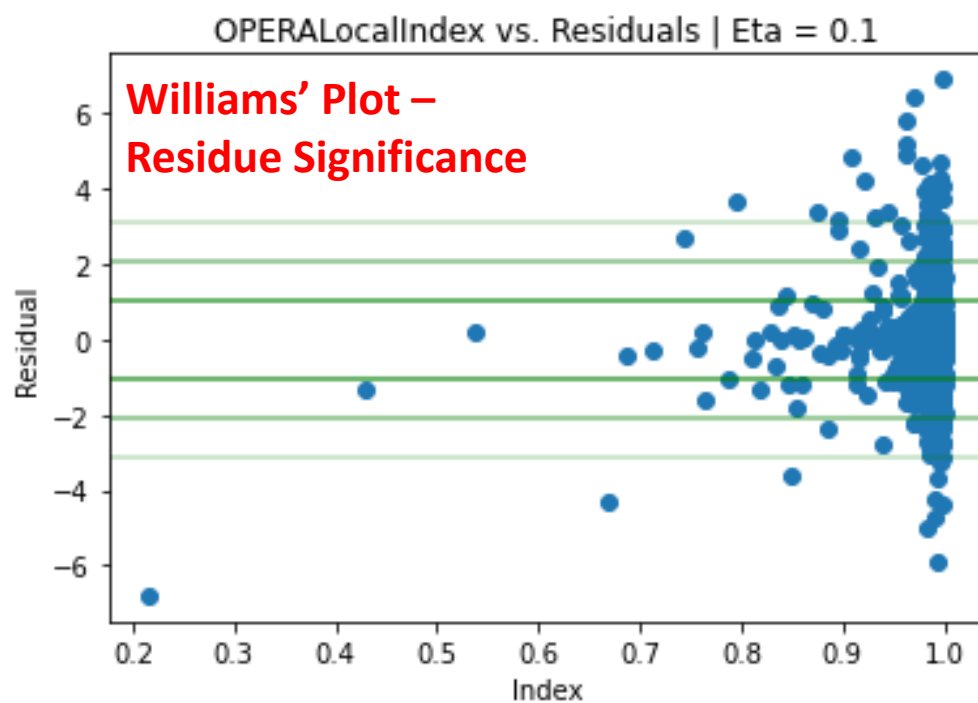
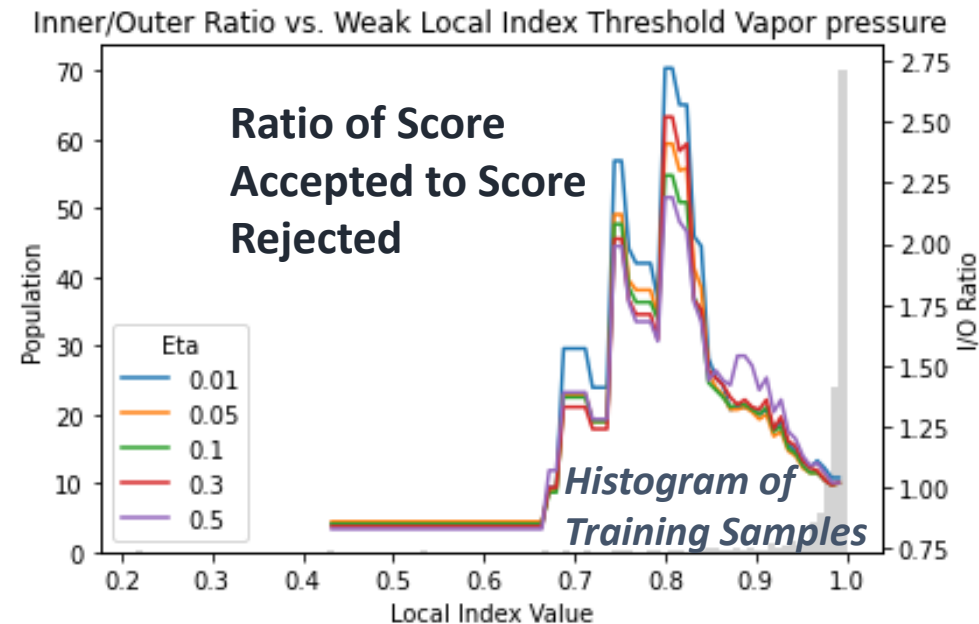
- Regression Architecture
 - We consider Extreme Gradient Boosted Trees regression with differing degrees of 'eta', the learning rate
 - 'eta' is a parameter that effects how much the algorithm can learn the fine details of the data, effecting how conservative the model is





Vapor pressure
Medium-Clean-Single Mechanism

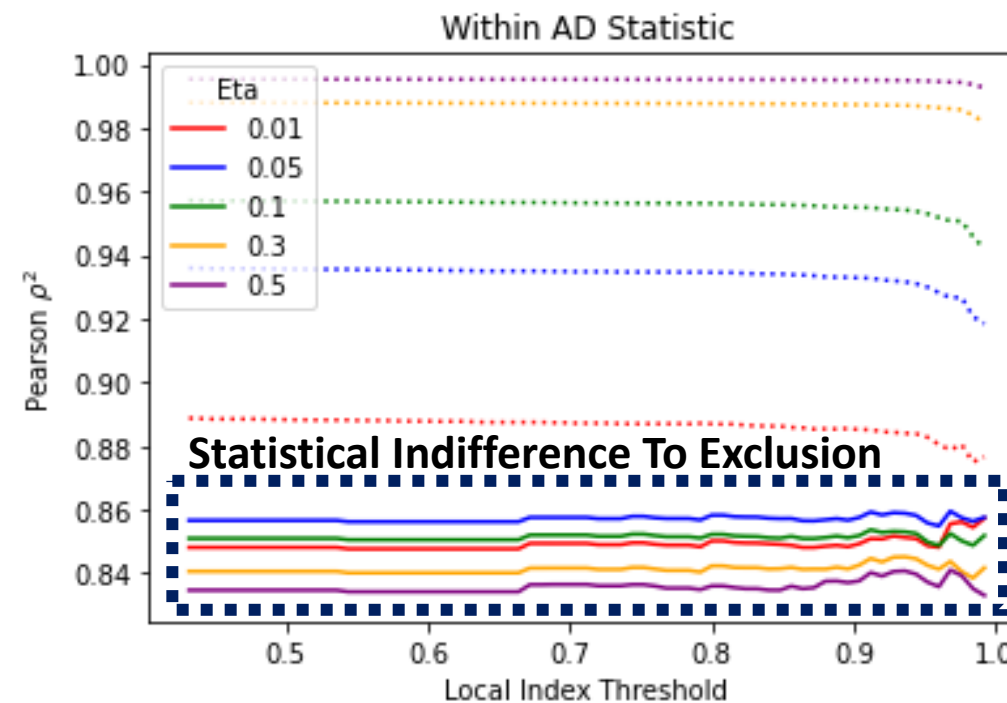
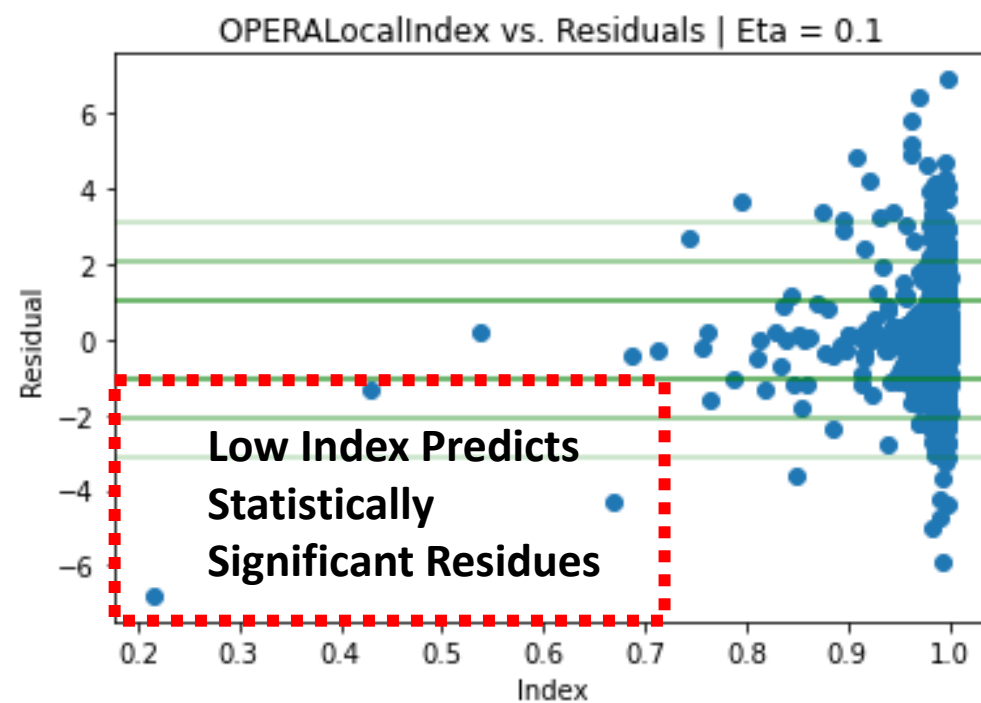
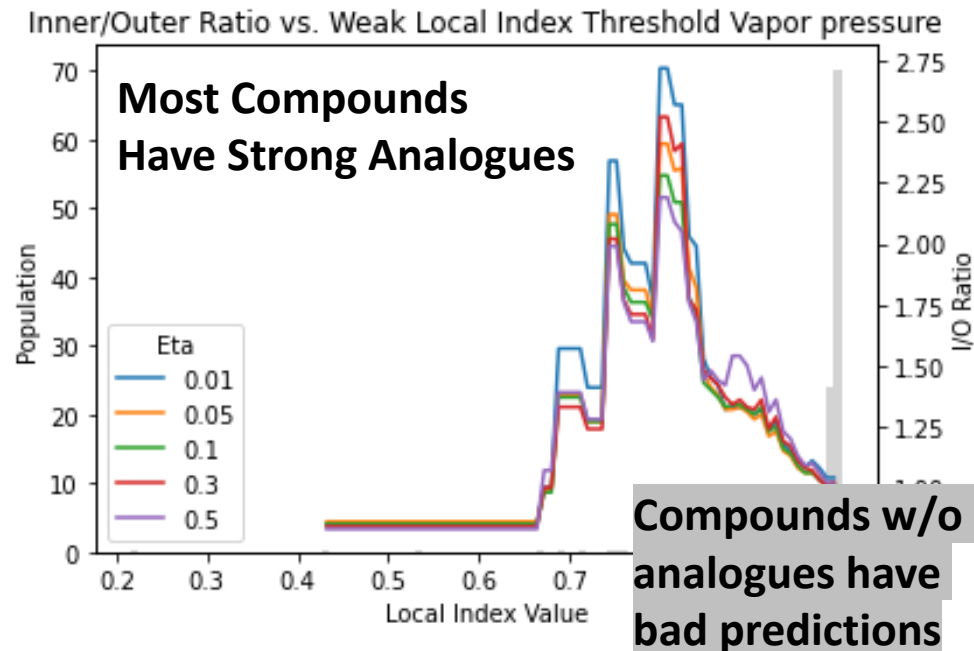
Vapor pressure





Vapor pressure
Medium-Clean-Single Mechanism
Training: ~1900 Testing: ~800

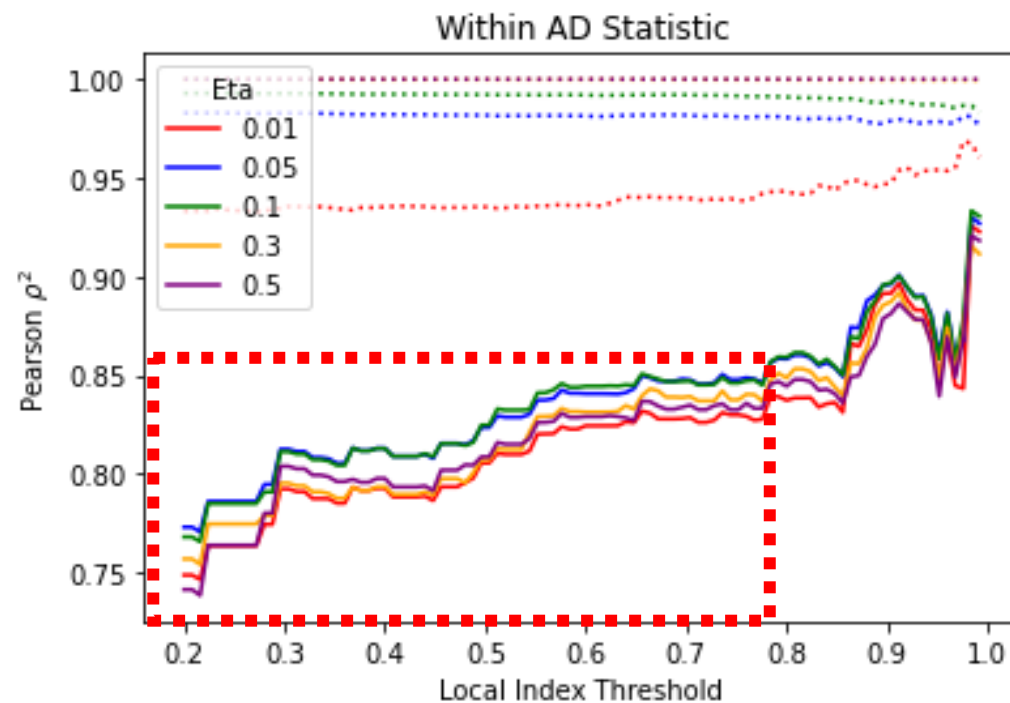
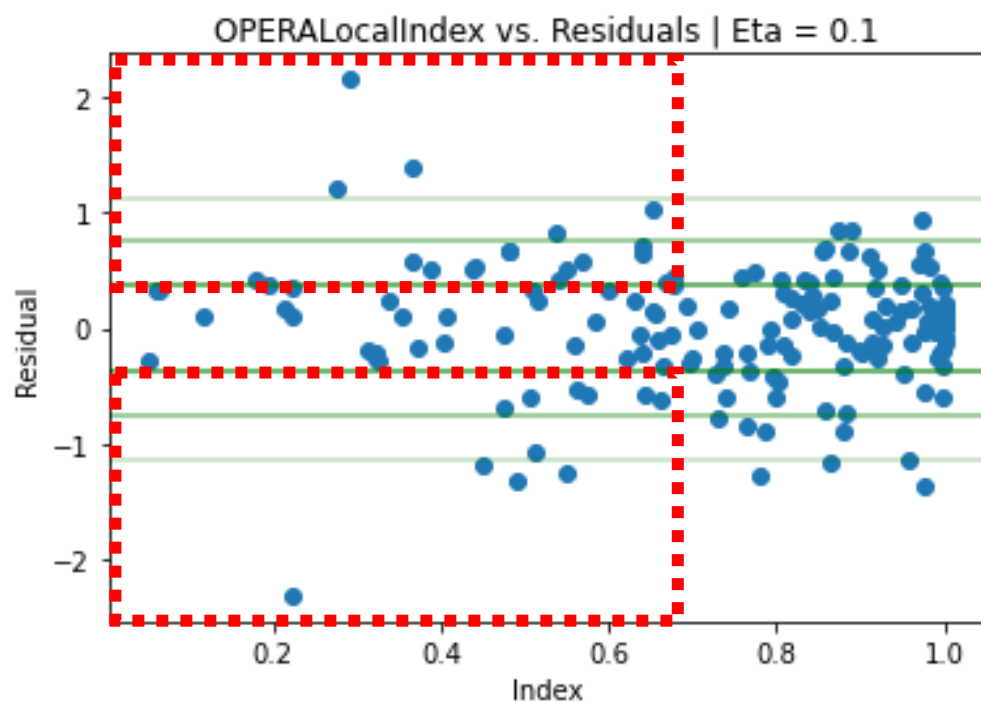
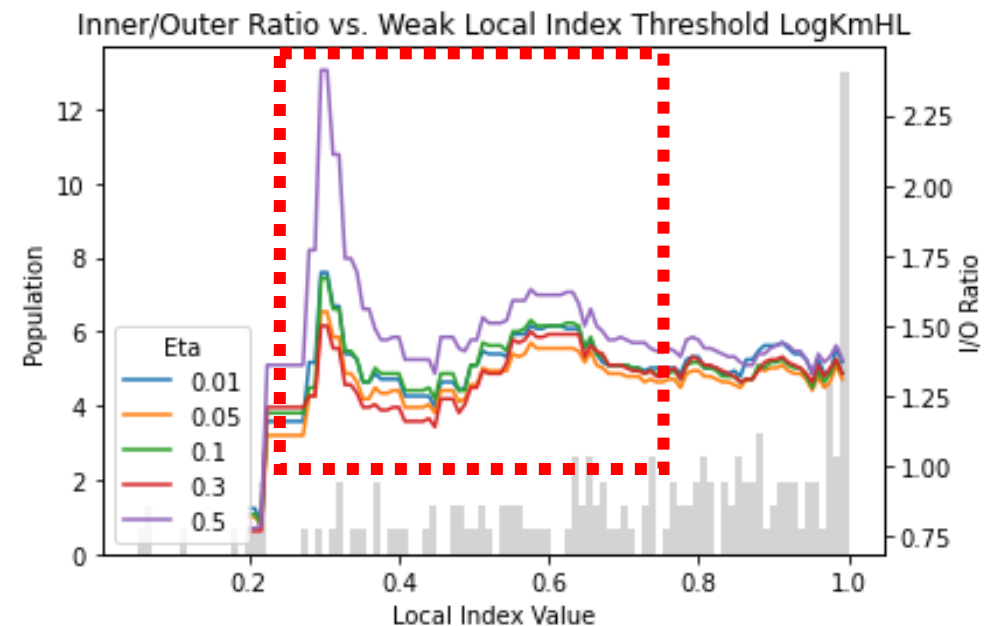
Vapor pressure



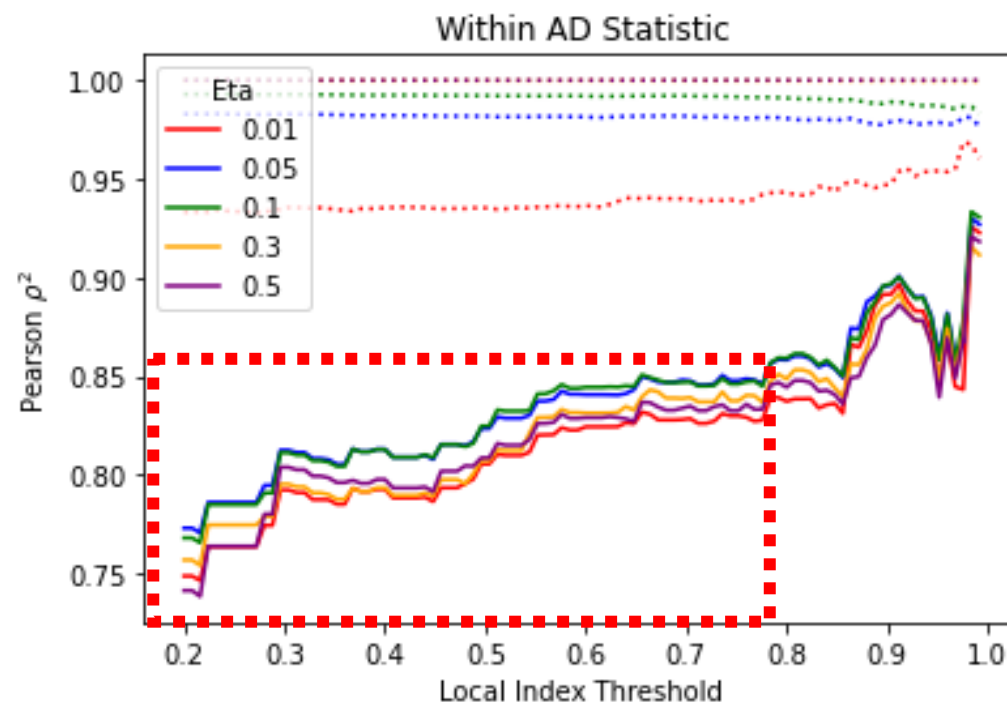
.....

Biotransformation by zebra fish
Small-Noisy-Multimechanistic

LogKmHL

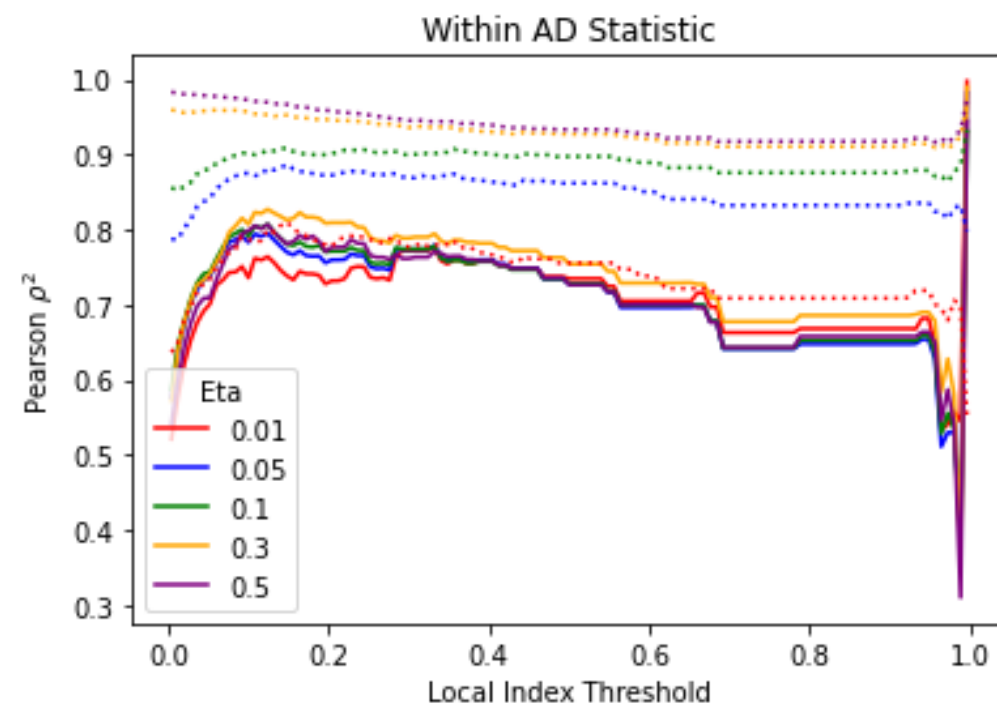
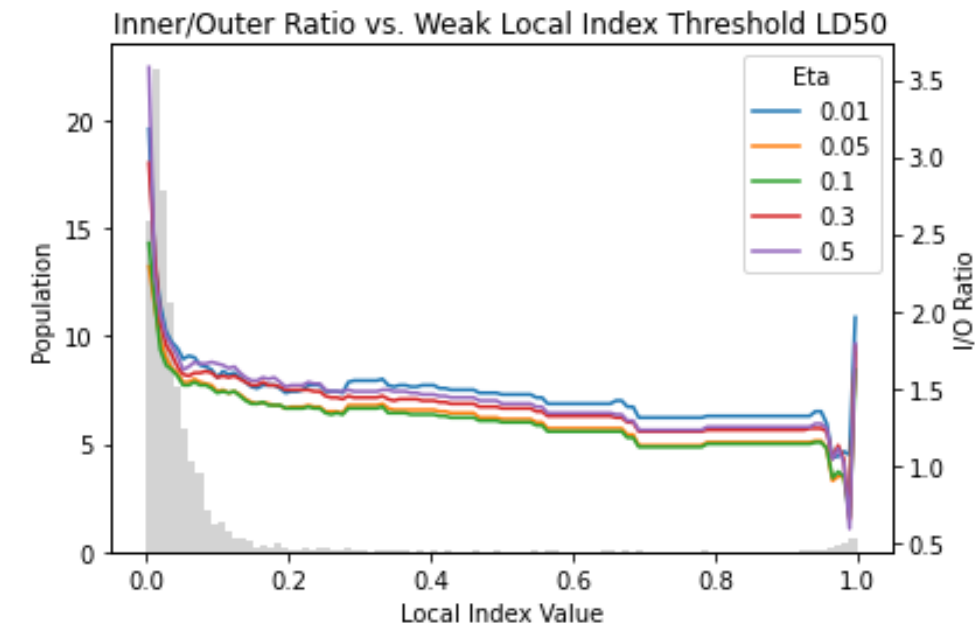
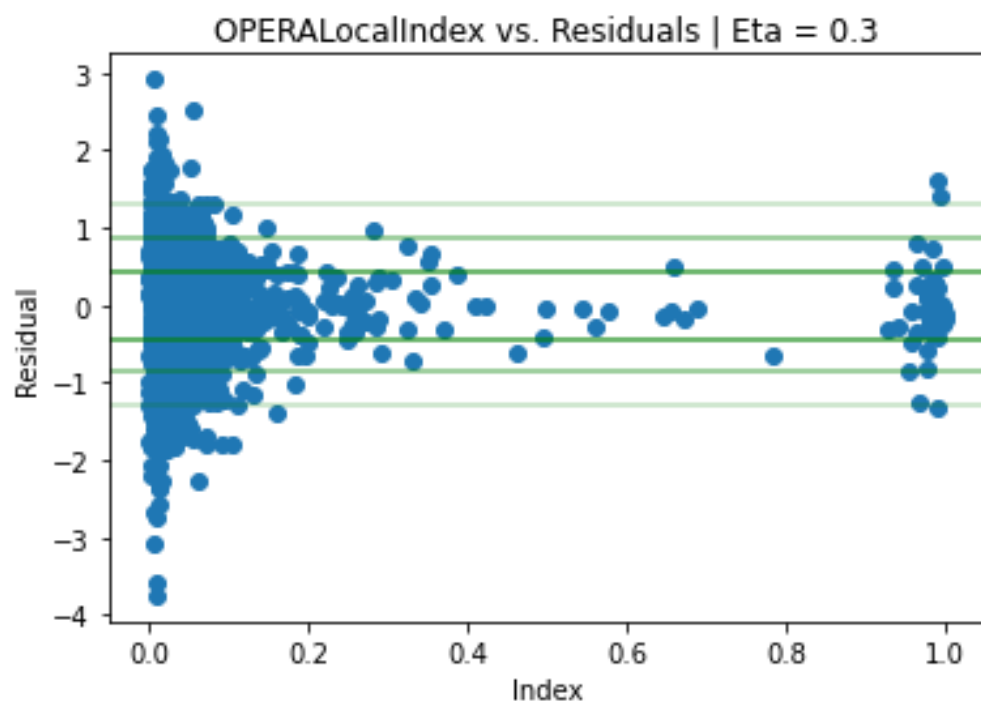


LogKmHL



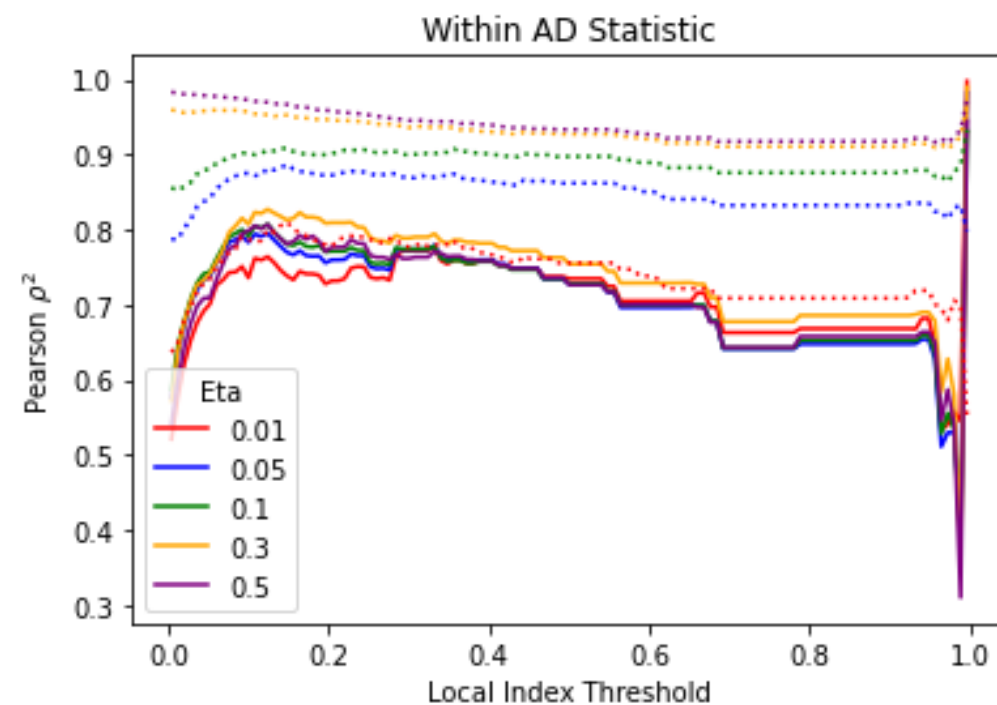
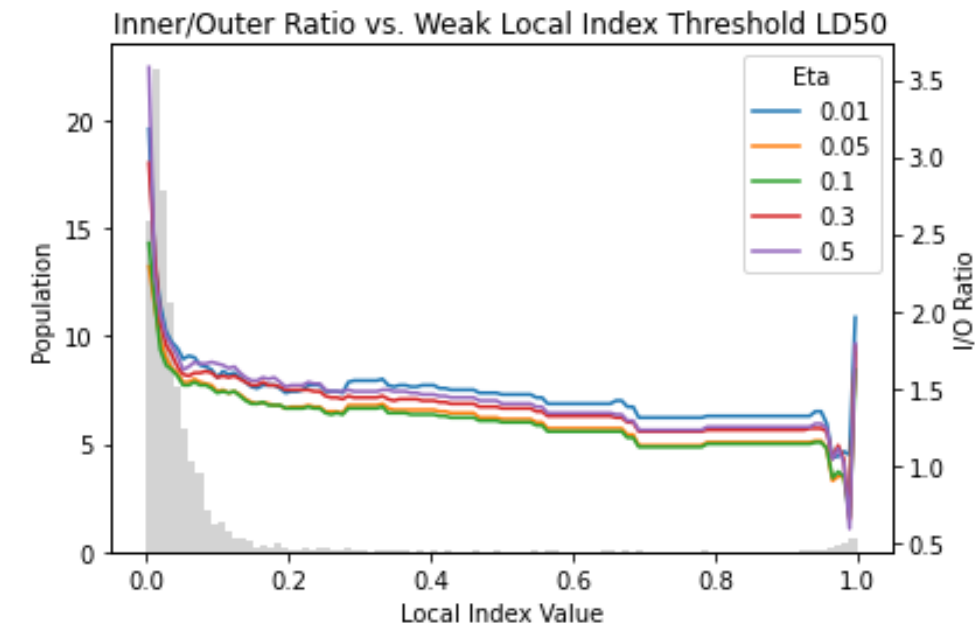
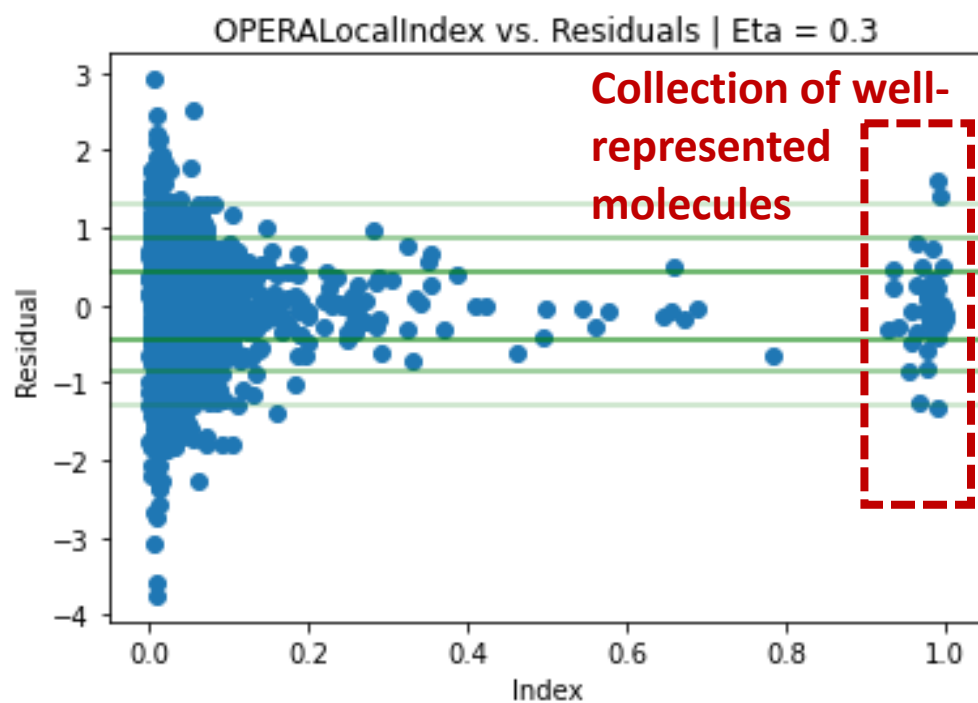
Oral lethal dose in rats
Big-Noisy-Multimechanistic

LD50



Oral lethal dose in rats
Big-Noisy-Multimechanistic
Training: ~5200 Testing: ~2200
Length of Embedding:

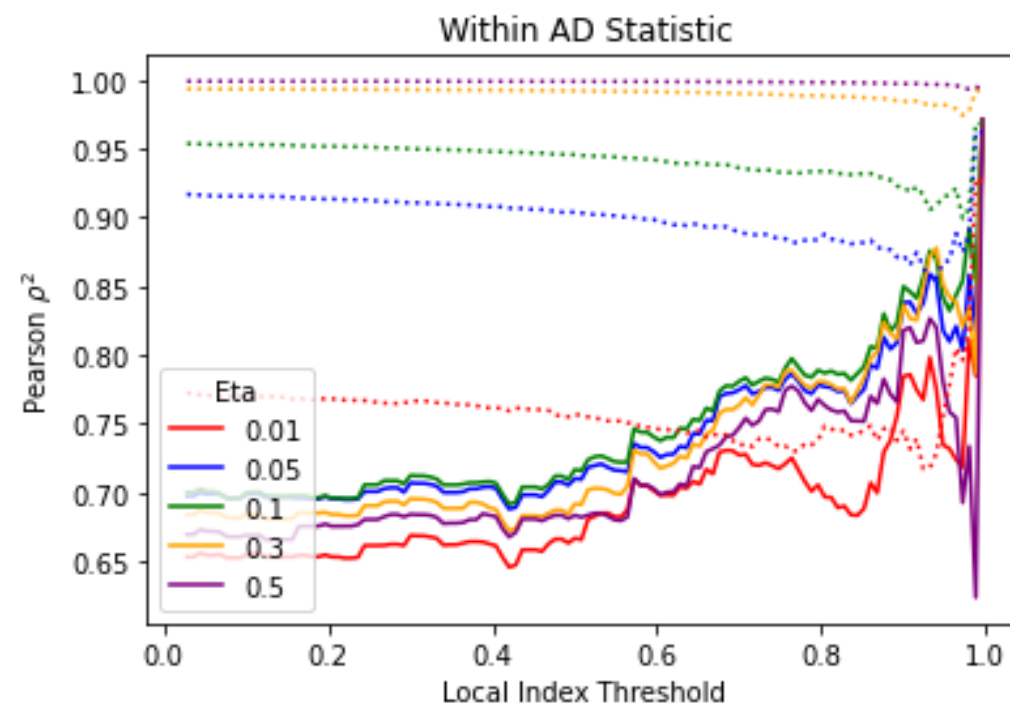
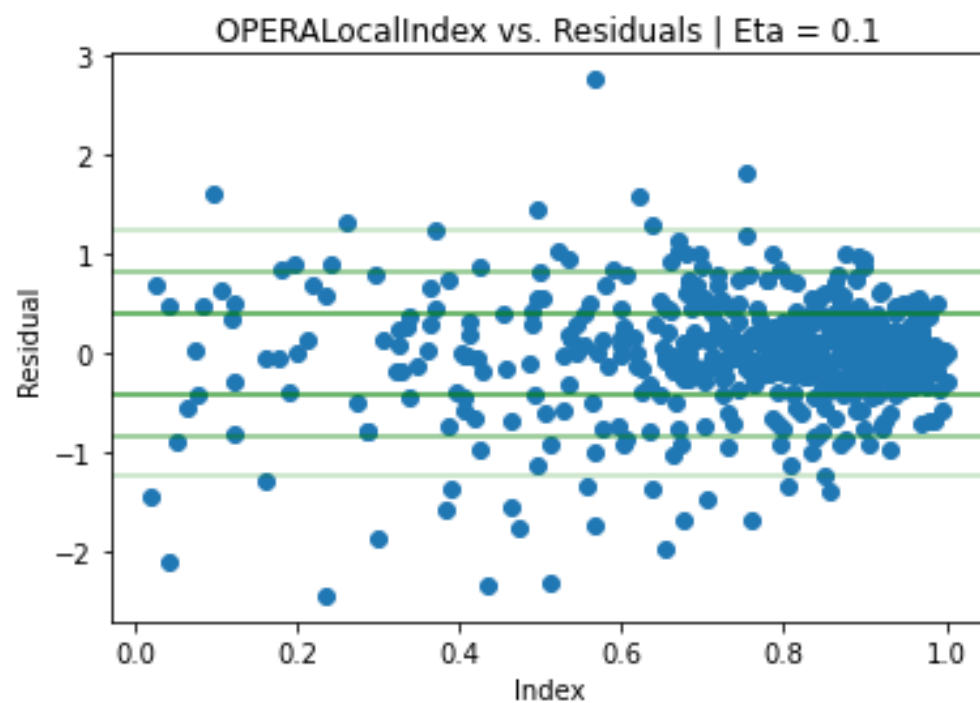
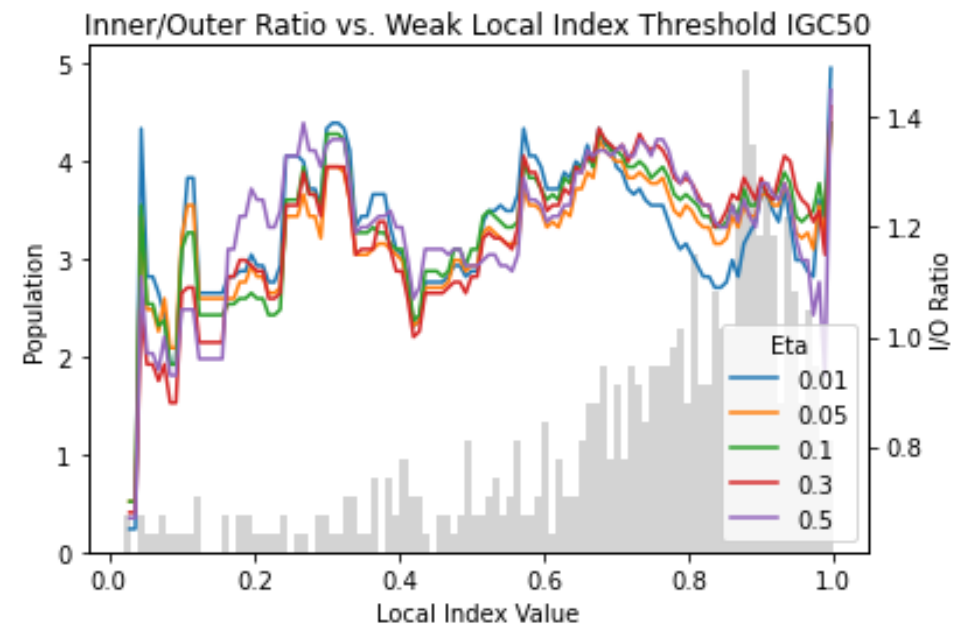
LD50





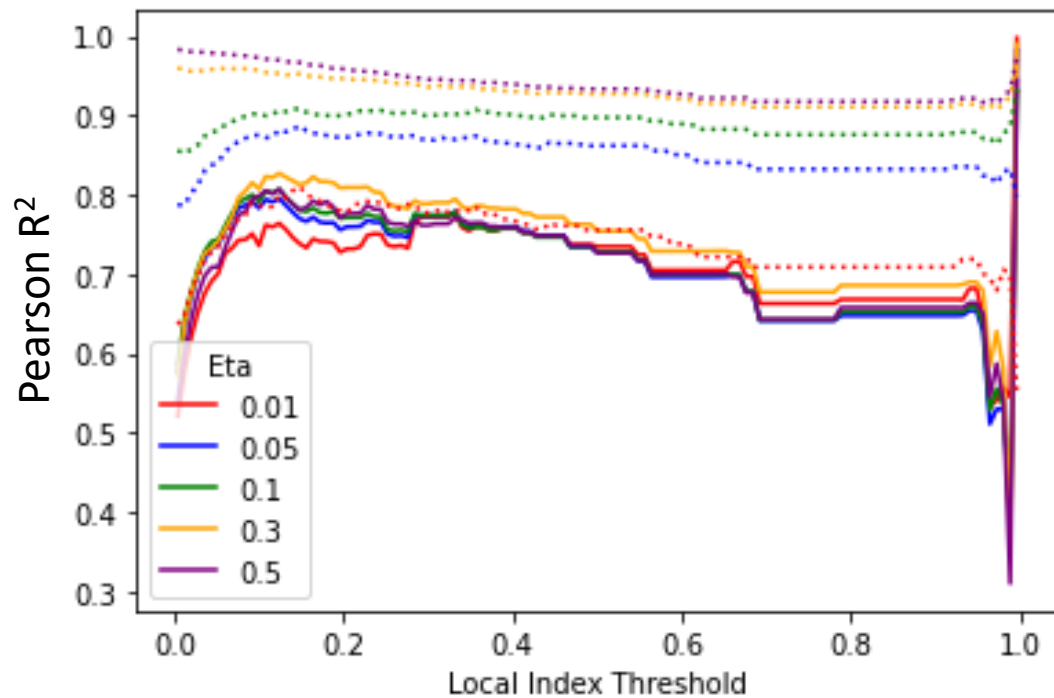
Inhibitory growth concentration
Medium-Noisy-Multimechanistic
Training: ~1250 Testing: ~540
Length of Embedding:

IGC50

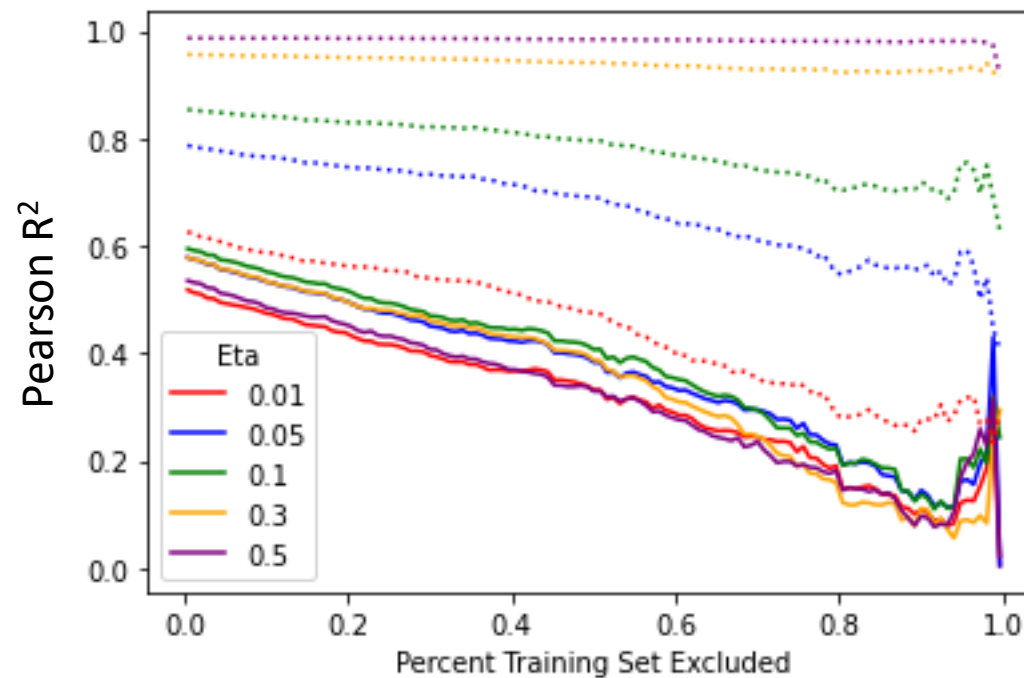


LD50 Comparison

OPERA Local Index

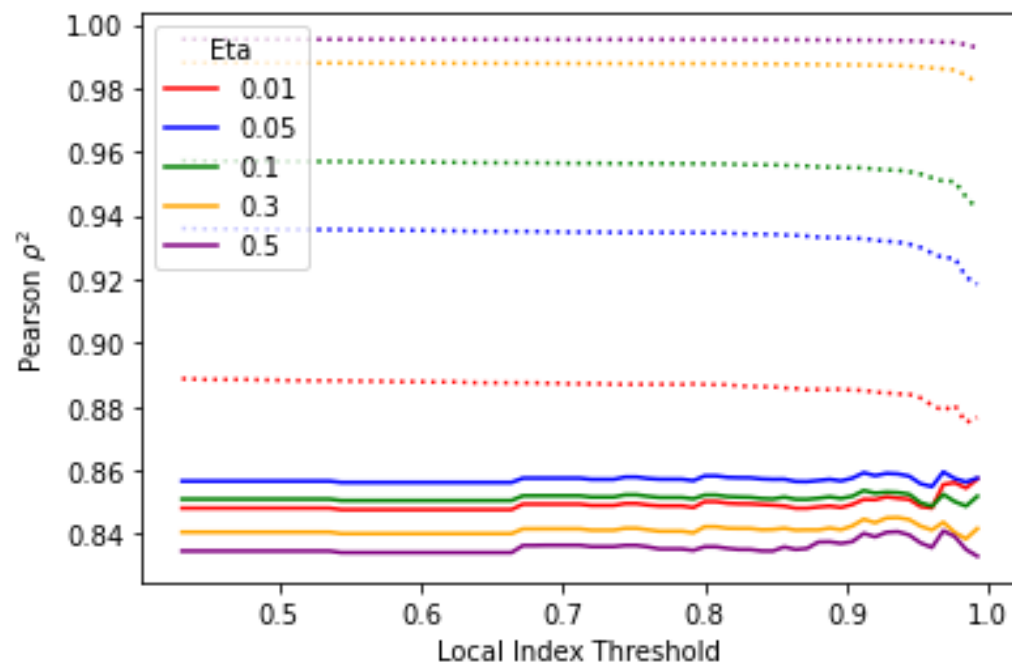


Average Cosine Similarity

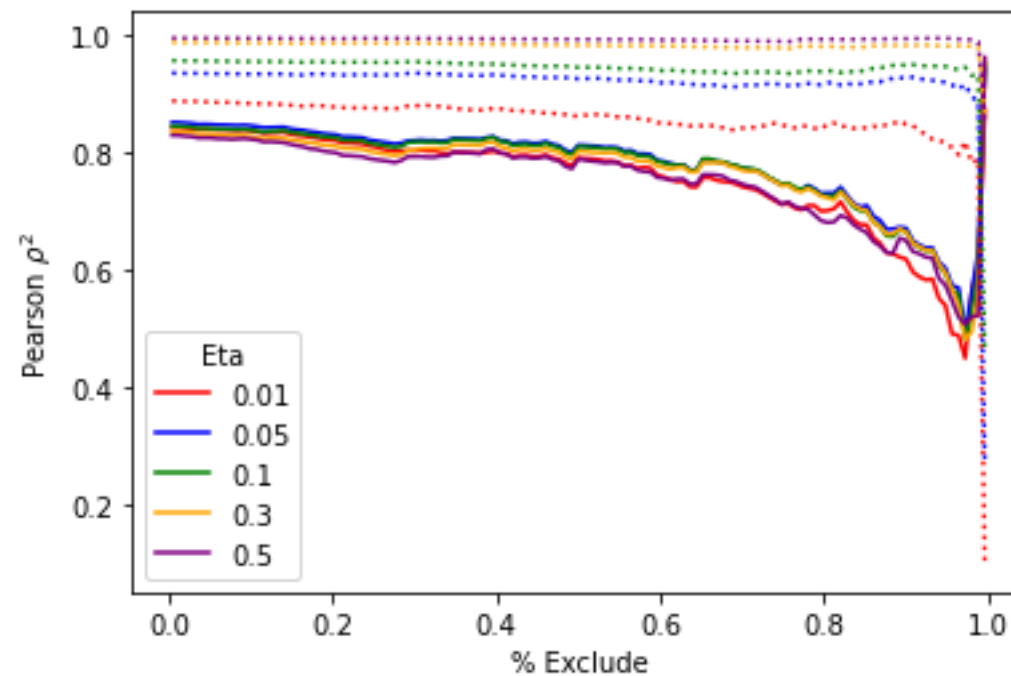


Vapor Pressure Comparison

OPERA Local Index

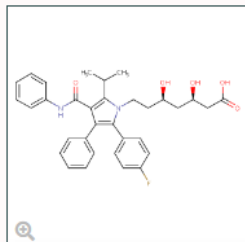


Average Cosine Similarity



OPERA Model Calculation Details: Water Solubility

[PRINT PDF](#)



Model Results

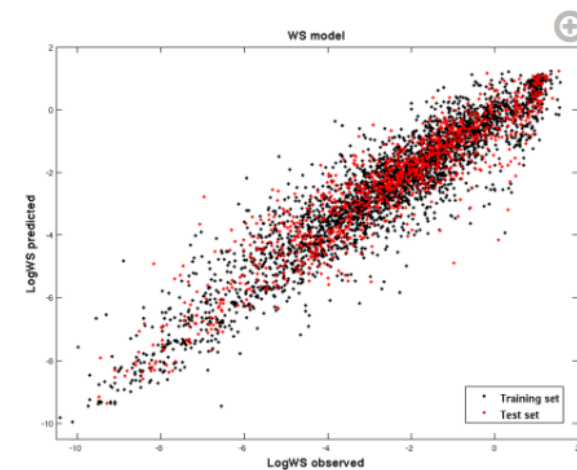
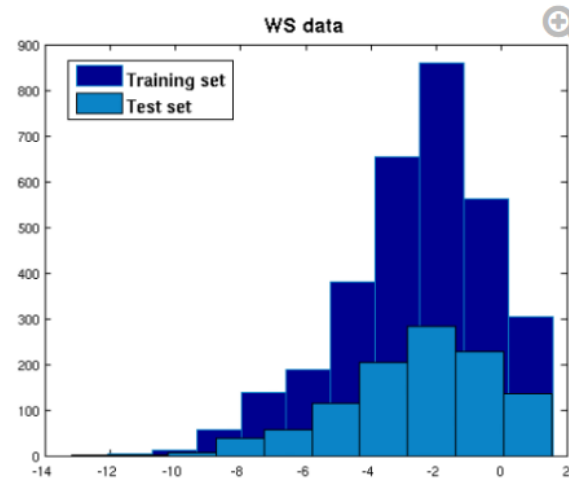
Predicted value: 6.73e-7 mol/L

Global applicability domain: Inside

Local applicability domain index: 0.260

Confidence level: 0.315

Model Performance



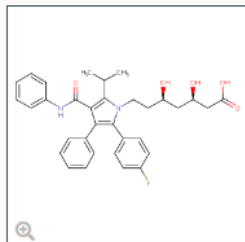
[QMRF](#)

Weighted KNN model

5-fold CV (75%)		Training (75%)		Test (25%)	
Q2	RMSE	RMSE	R2	RMSE	R2
0.870	0.810	0.820	0.870	0.860	0.860

OPERA Model Calculation Details: Water Solubility

PRINT PDF



Model Results

Predicted value: 6.73e-7 mol/L

Global applicability domain: Inside

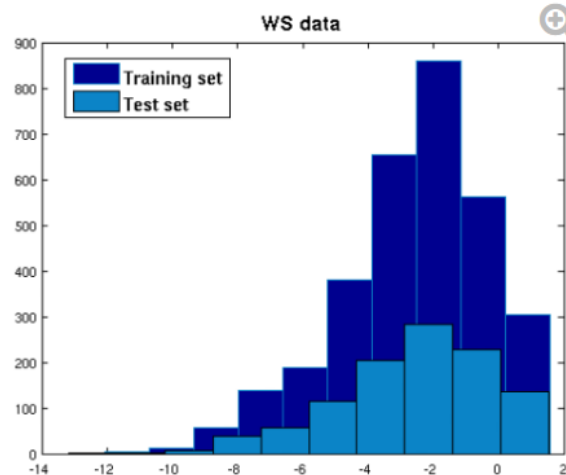
Local applicability domain index: 0.260

Confidence level: 0.315

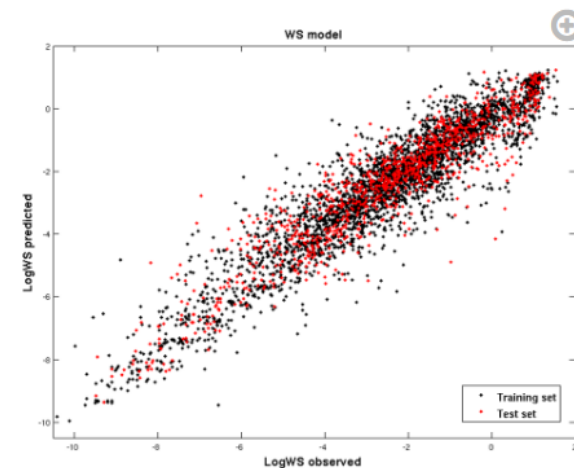
Explore prediction confidence

Expert Annotations

Model Performance



Applicability Domain Graphics?



QMRF

Weighted KNN model

5-fold CV (75%)		Training (75%)		Test (25%)	
Q2	RMSE	RMSE	R2	RMSE	R2
0.870	0.810	0.820	0.870	0.860	0.860

Conclusions

- Because of the vast diversity of mathematical relationships driving SARs, it is unlikely any *one* method will be suitable for every use case
 - We must be cautious in ensuring our models consider things like the size of data, the diversity of data, and the complexity of the endpoint

Conclusions

- Because of the vast diversity of mathematical relationships driving SARs, it is unlikely any *one* method will be suitable for every use case
 - We must be cautious in ensuring our models consider things like the size of data, the diversity of data, and the complexity of the endpoint
- The future of applicability domain may be in confidence interval approaches or more detailed leveraging of machine learning to characterize the *shape* of the data patterns within the model as part of work-up

Conclusions

- Because of the vast diversity of mathematical relationships driving SARs, it is unlikely any *one* method will be suitable for every use case
 - We must be cautious in ensuring our models consider things like the size of data, the diversity of data, and the complexity of the endpoint
- The future of applicability domain may be in confidence interval approaches or more detailed leveraging of machine learning to characterize the *shape* of the data patterns within the model
- OPERA pursues the “confidence index”, and EPA is placing effort into rating predictions in addition to defining applicability domain