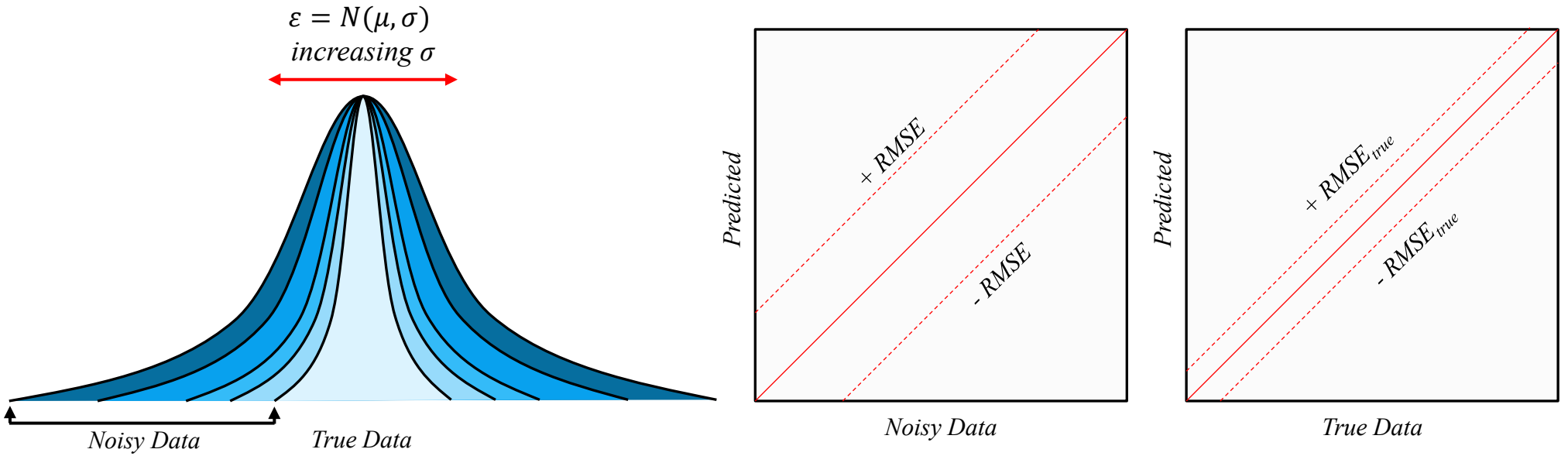
Determining the Predictive Limit of QSAR Models



Scott Kolmar

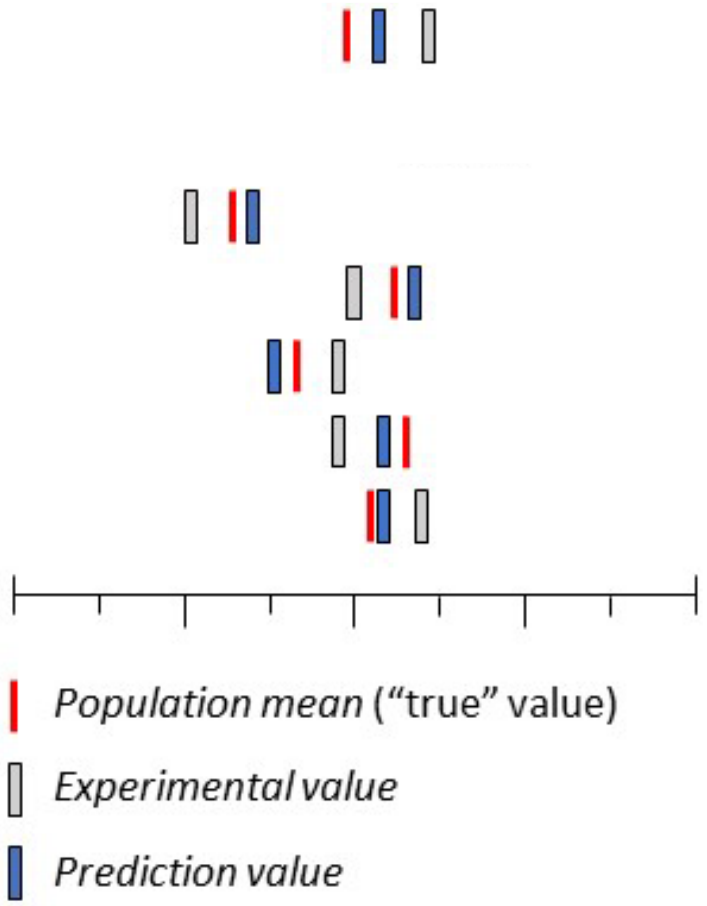
ORCID: 0000-0002-7797-700X, Kolmar.Scott@epa.gov

June 24th, 2021

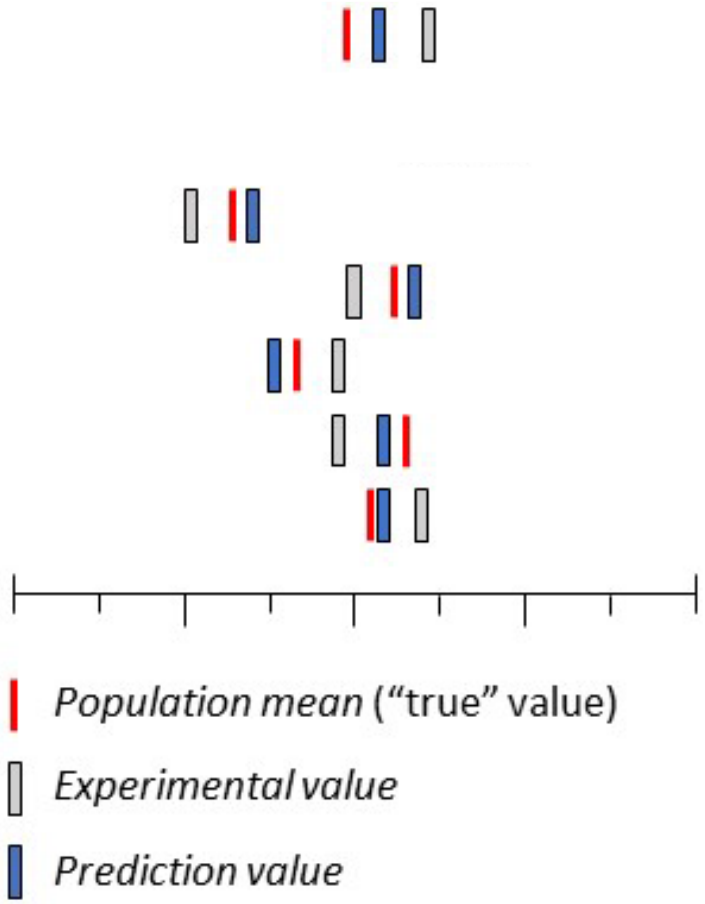
US EPA

ORD-CCTE-CCED-CCCB

Evaluating QSAR Models

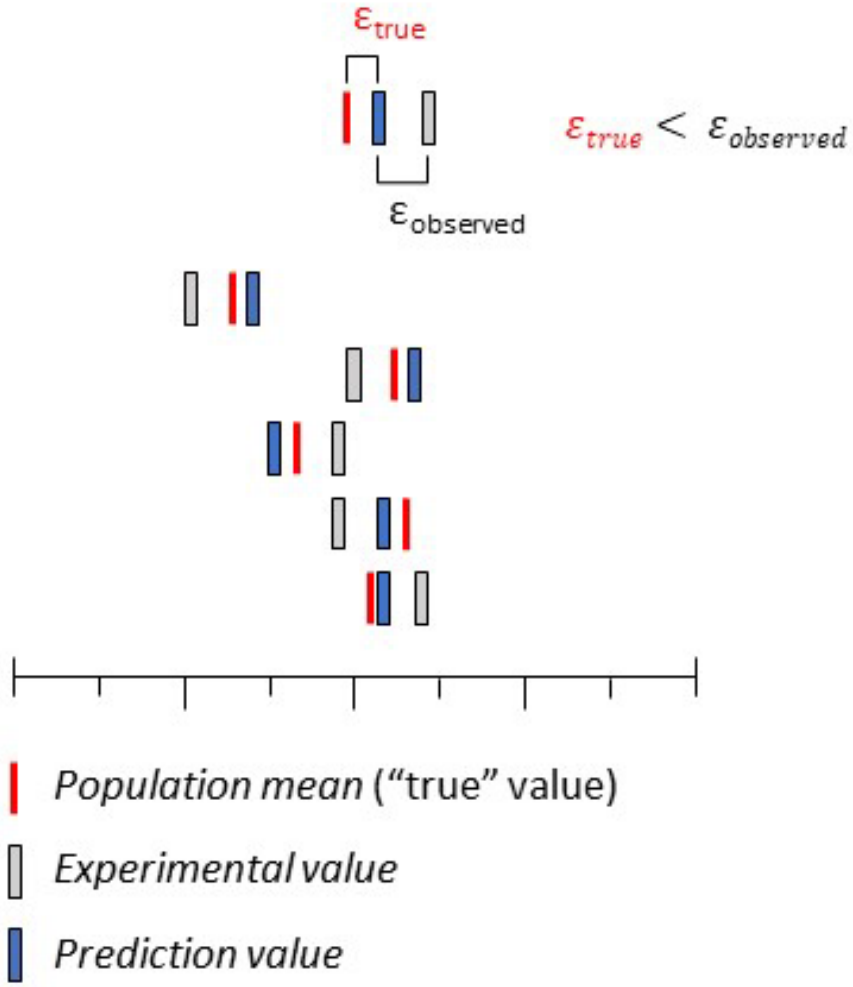


Evaluating QSAR Models



QSAR models attempt to predict the *population mean*

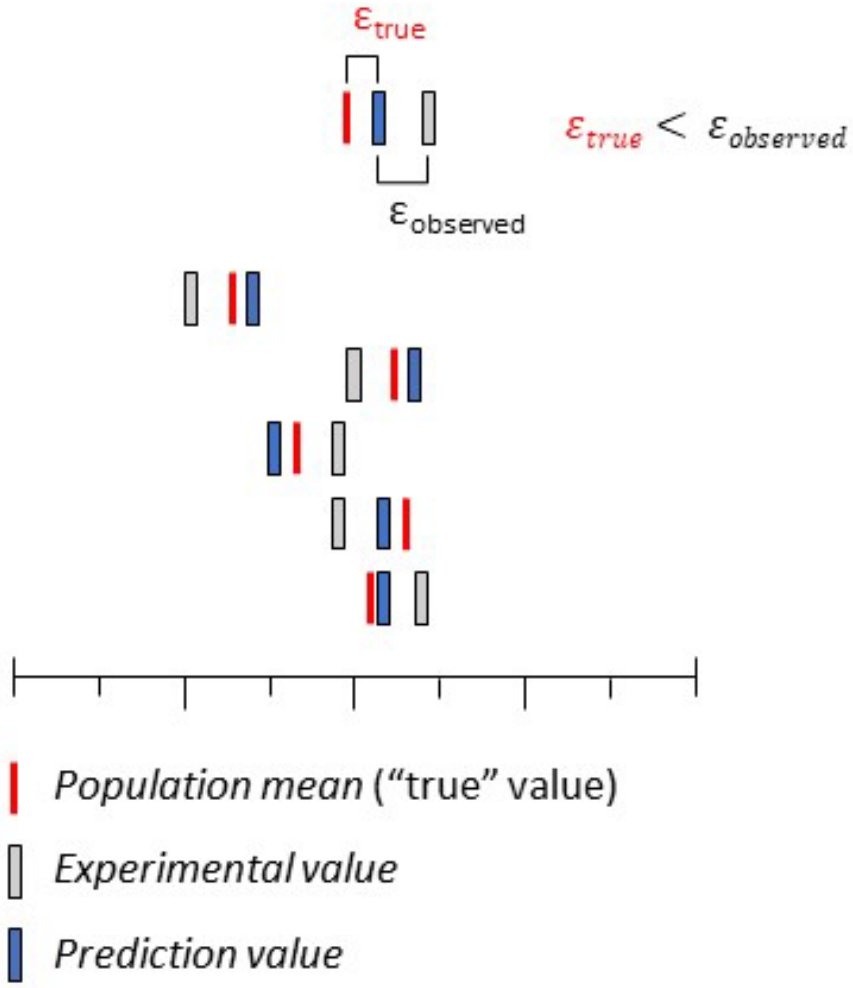
Evaluating QSAR Models



QSAR models attempt to predict the *population mean*

QSAR models are evaluated by $\varepsilon_{observed}$

Evaluating QSAR Models

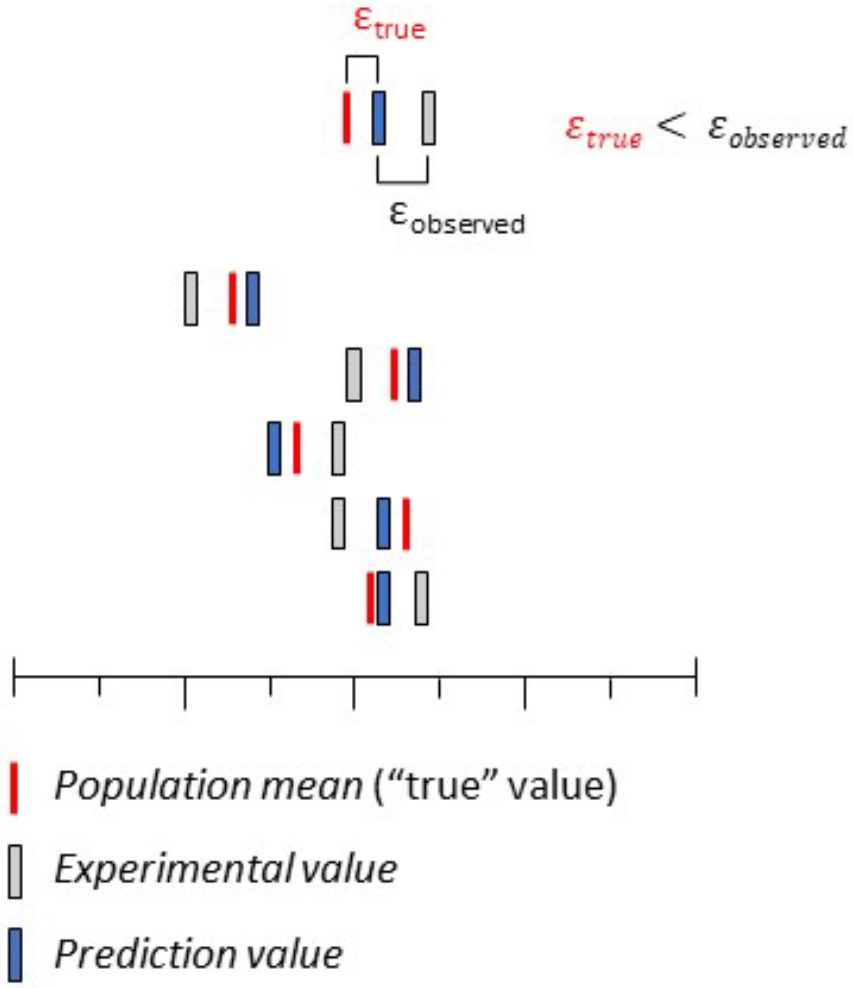


QSAR models attempt to predict the *population mean*

QSAR models are evaluated by $\varepsilon_{observed}$

This evaluation is flawed because the *experimental value* is not overlapping with the *population mean*

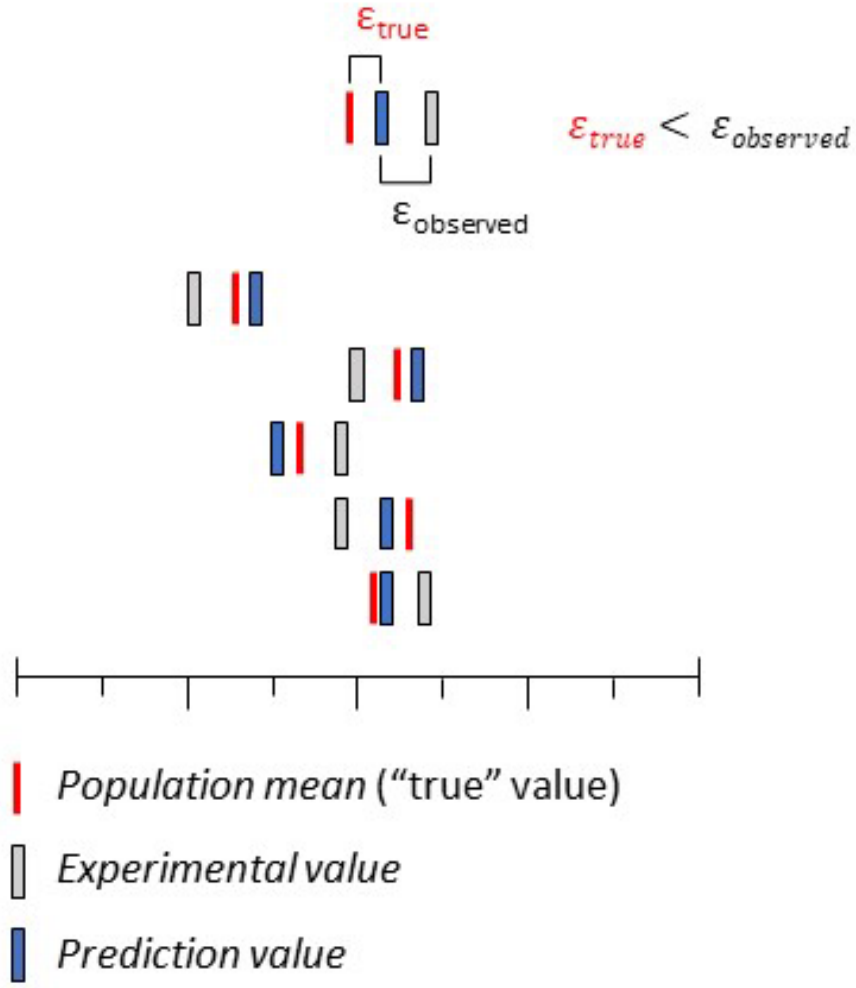
Evaluating QSAR Models



Research Question

Population means are difficult to measure or are generally unavailable in typical QSAR datasets. How can we judge the quality of a QSAR model when it is trained on *experimental values* which do not represent *population means*?

Evaluating QSAR Models



Research Question

Population means are difficult to measure or are generally unavailable in typical QSAR datasets. How can we judge the quality of a QSAR model when it is trained on *experimental values* which do not represent *population means*?

Strategy

- Designate the *experimental values* as "*population means*"
- Add simulated error to these values
- Train a QSAR model on the error laden values
- Make predictions
- Evaluate if the predictions are closer to our designated "*population means*" or the error laden values

Experimental Error in QSAR

response variable	number of molecules	number of results	number of molecules to consider	percentage of data set with a single measurement
human hep CL_{int}	10668	22588	9819	40
human mic CL_{int}	32492	47566	31215	74
human PPB	61356	80725	59852	89
log $D_{7.4}$	115441	140662	113339	93
rat hep CL_{int}	39112	55969	36807	77
rat PPB	16476	23738	16037	85
solubility (dried DMSO)	44256	49043	42821	95
solubility (solid)	38722	42736	36256	95

Wenlock et al. *J. Chem. Inf. Model.*, **2015**, 55, 125

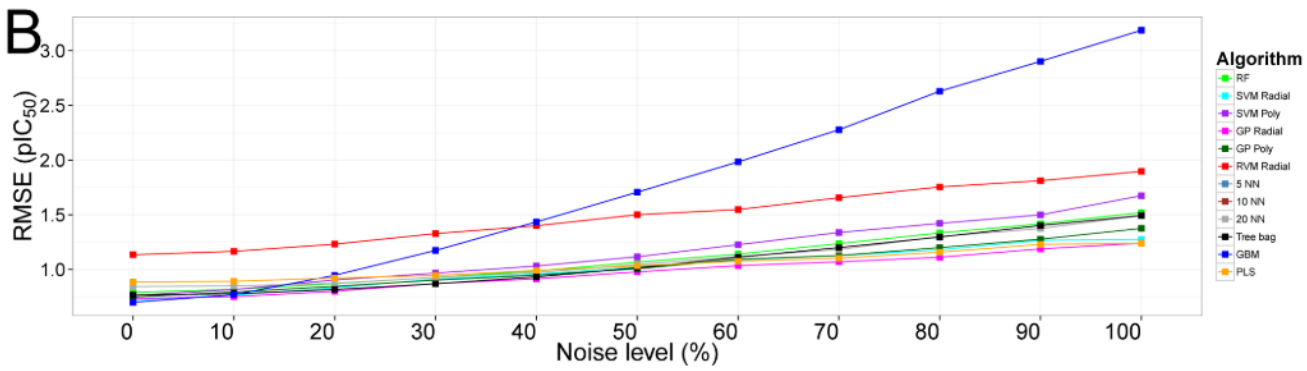
- Uncertainty information from multiple measurements is rare in cheminformatics

Experimental Error in QSAR

response variable	number of molecules	number of results	number of molecules to consider	percentage of data set with a single measurement
human hep CL_{int}	10668	22588	9819	40
human mic CL_{int}	32492	47566	31215	74
human PPB	61356	80725	59852	89
log $D_{7.4}$	115441	140662	113339	93
rat hep CL_{int}	39112	55969	36807	77
rat PPB	16476	23738	16037	85
solubility (dried DMSO)	44256	49043	42821	95
solubility (solid)	38722	42736	36256	95

Wenlock et al. *J. Chem. Inf. Model.*, **2015**, 55, 125

- Uncertainty information from multiple measurements is rare in cheminformatics



Cortes-Ciriano et al. *J. Chem. Inf. Model.*, **2015**, 55, 1413

- Simulated error can elicit different responses from different algorithms; certain hyperparameters govern these responses

Error in QSAR

Explicit Assumption: Predictions must have uncertainty higher than or equal to the training set.

"It follows that the model's prediction of the *external test set* will have uncertainty equal to or greater than that contained within the *training set*."

Wenlock et al. *J. Chem. Inf. Model.*, **2015**, 55, 125

"The experimental uncertainty sets the *upper limit of performance* of in silico models that can be achieved."

Kramer et al. *J. Med. Chem.,* **2012**, 55, 5165

Error in QSAR

Explicit Assumption: Predictions must have uncertainty higher than or equal to the training set.

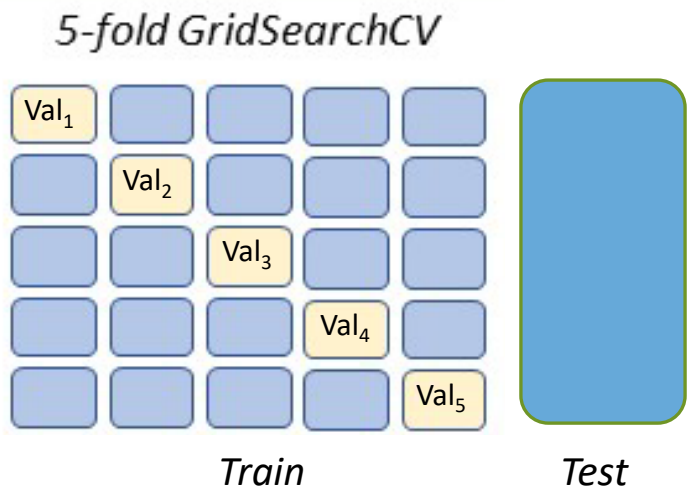
"It follows that the model's prediction of the *external test set* will have uncertainty equal to or greater than that contained within the *training set*."

Wenlock et al. *J. Chem. Inf. Model.*, **2015**, 55, 125

"The experimental uncertainty sets the *upper limit of performance* of in silico models that can be achieved."

Kramer et al. *J. Med. Chem.*, **2012**, 55, 5165

Implicit Assumption: Test sets do not have error and can be used to evaluate QSAR models.



- *Train* is commonly acknowledged to contain error
- It is assumed that *Test* has no error
- Models are evaluated on their ability to predict *error laden* data

Error in QSAR

This work seeks to directly test the hypothesis
that a model's *prediction uncertainty* is limited by the *uncertainty in the training data*

Datasets:

- Span a range of complexity from quantum mechanical to *in vivo* toxicological
- Represent endpoints of interest in QSAR
- The series of datasets will have endpoints with increasing levels of experimental uncertainty

Methods:

- Add simulated error to each dataset
- Build models on the *error laden data*
- Make predictions
- Evaluate predictions against the *true values*
- Evaluate predictions against the *error laden values*
- Compare model performance

Datasets

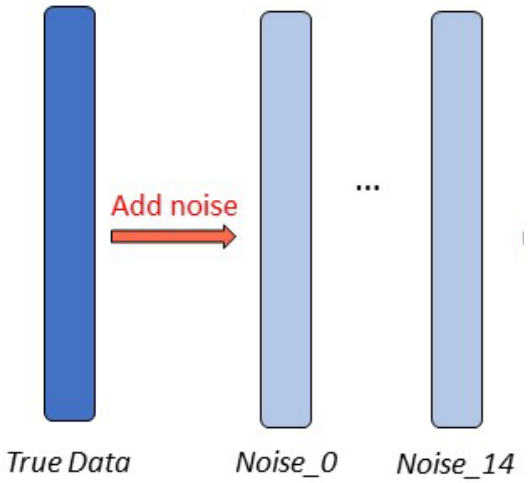
Dataset	Category	Number of Molecules[a]	Endpoint	Range
G298_atom	Quantum Mechanical	131,082	ΔG^{o}_{at} (kcal mol^{-1})	-2,417 – -288
Alpha	Quantum Mechanical	131,082	α (Bohr^{3})	9.0 – 27.8
Lip	Physiochemical	4,200	logD	-1.5 – 4.5
Solv	Physiochemical	642	ΔG^{o}_{hyd} (kcal mol^{-1})	-25.5 – 3.4
BACE	Biochemical	1,513	pIC_{50}	2.5 – 10.5
Tox_102[b]	Toxicological *in vitro*	971	$logAC_{50}$	-2.1 – 4.7
Tox_134[c]	Toxicological *in vitro*	1,347	$logAC_{50}$	-4.0 – 2.8
LD50	Toxicological *in vivo*	5,003	$logLD_{50}$ (mg kg^{-1})	-1.9 – 4.8

[a] Original size of the dataset. If datasets have more than 1,000 molecules, they were randomly sampled down to a size of 1,000 before modeling.

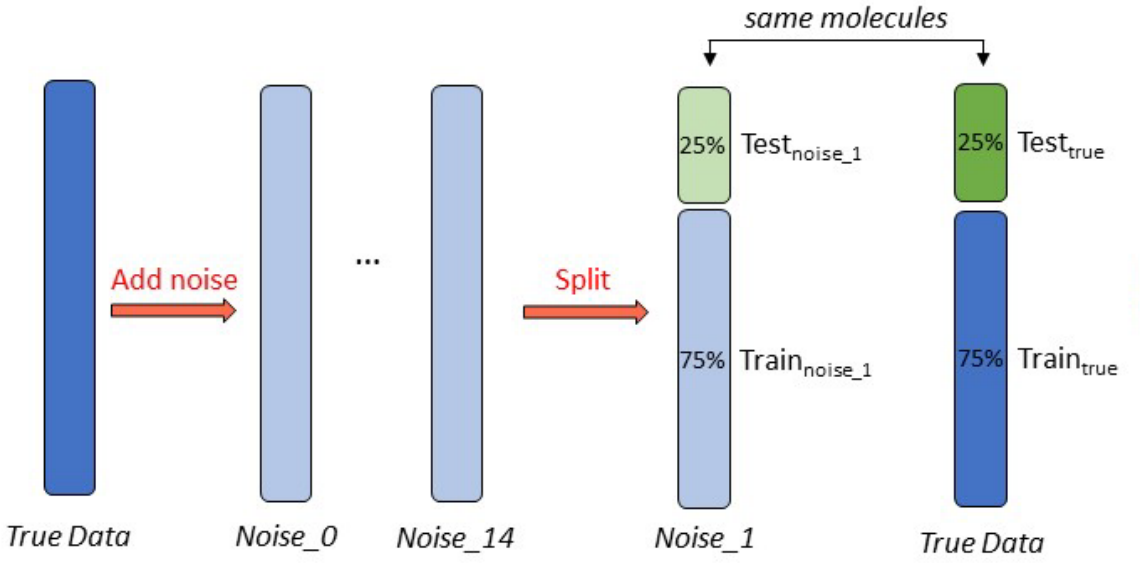
[b] Includes data exclusively from the ATG-PPre-cis assay

[c] Inclues data exclusively from the ATG-PPARg-trans assay

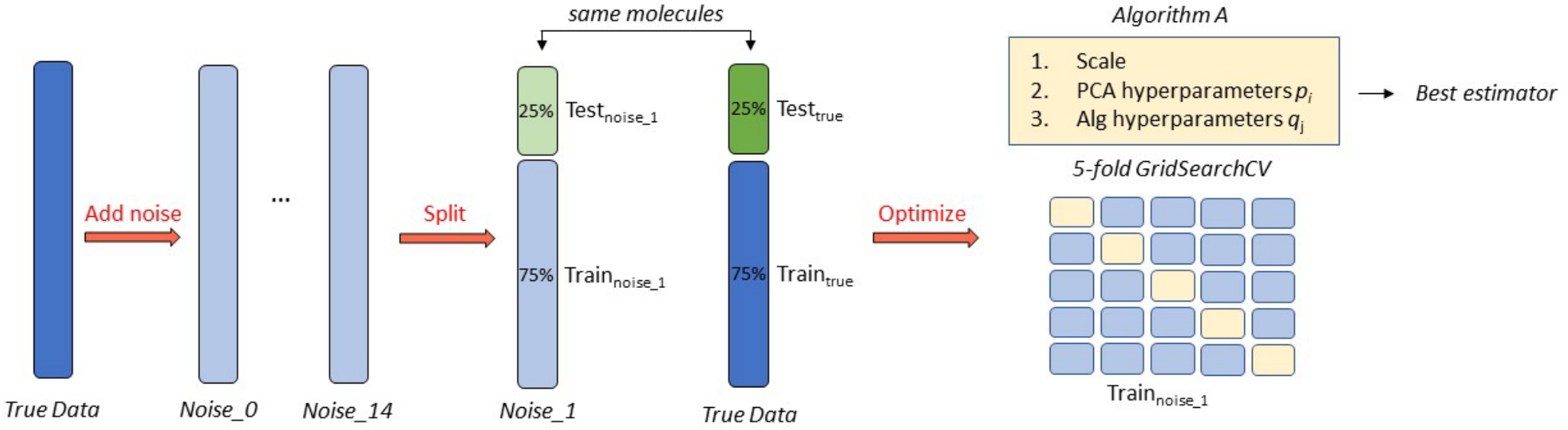
Modeling Workflow



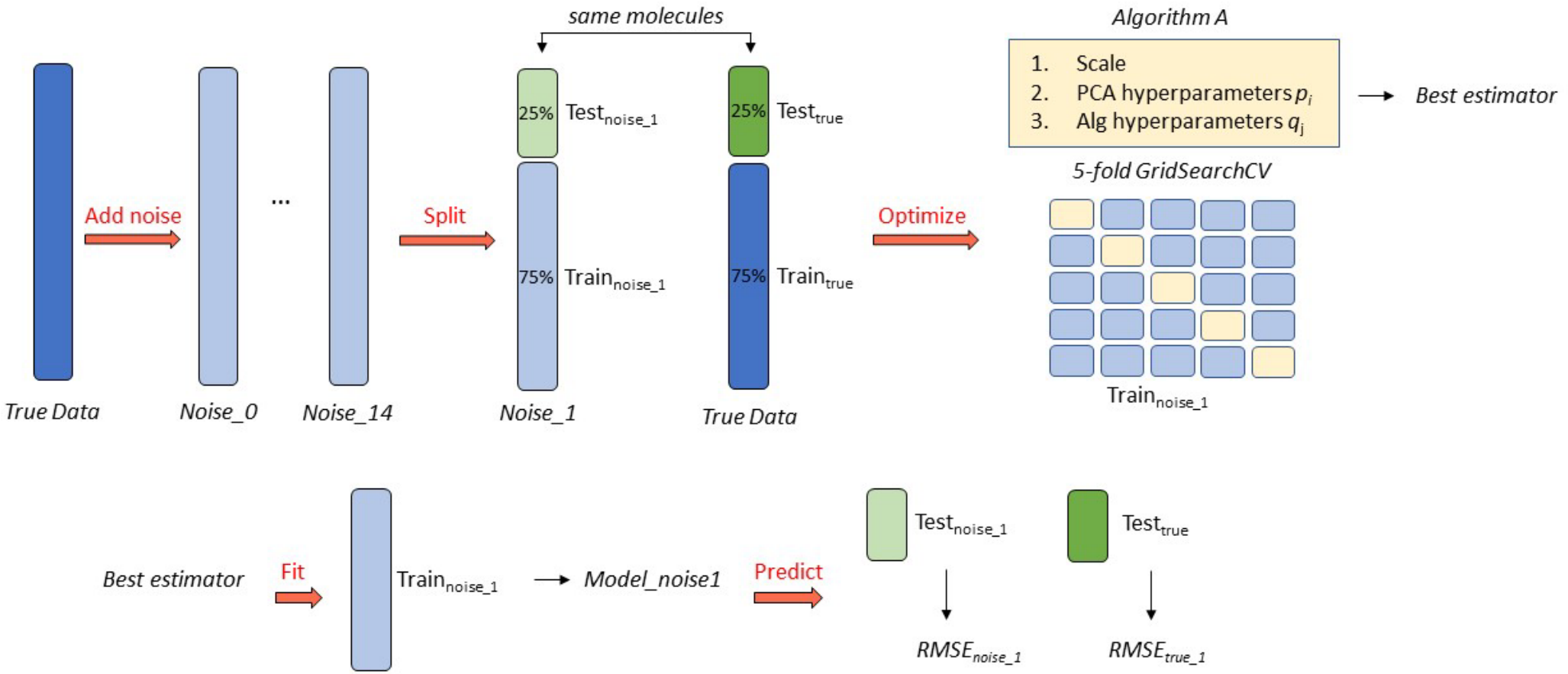
Modeling Workflow



Modeling Workflow



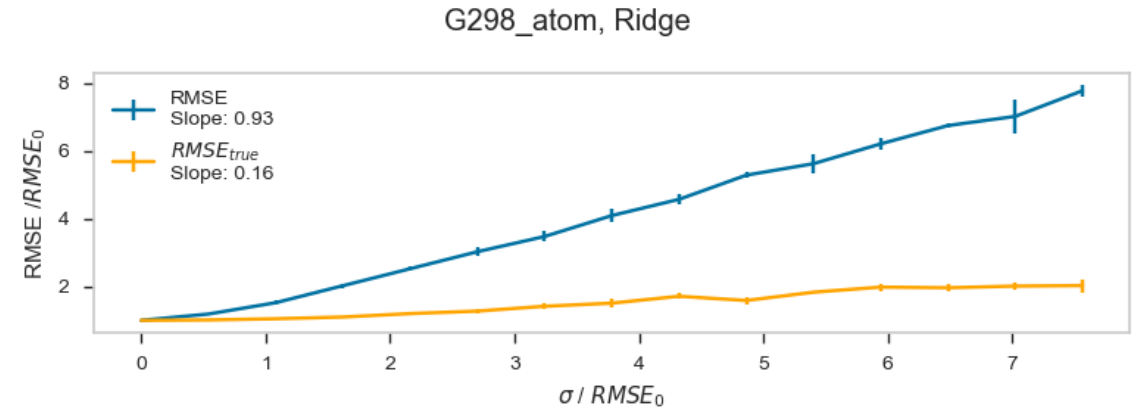
Modeling Workflow

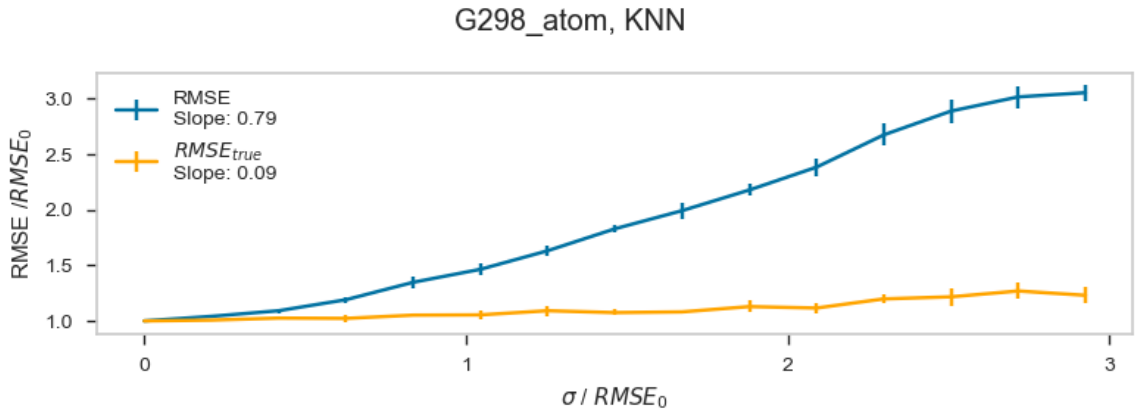


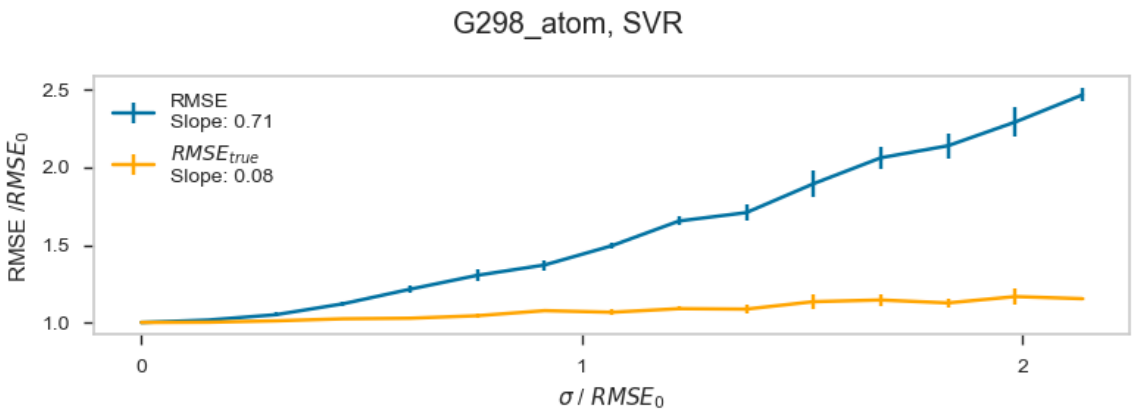
Algorithms and Hyperparameters

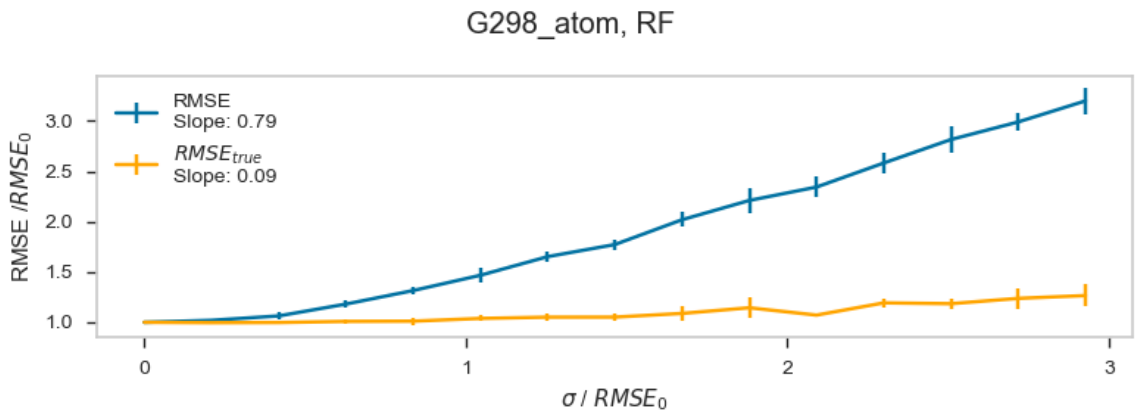
Algorithm	Hyperparameters Searched in Optimization
Ridge Regression (Ridge)	*PCA n components* $\in (1, 3, \ \dots\ , 59)$
	$\alpha \in (1, 2, 3, 4, 5, 10)$
k- Nearest Neighbors (kNN)	*PCA n components* $\in (1, 3, \ \dots\ , 59)$
	$k \in (1, 2, \dots, 20)$
Support Vector Regressor (SVR)	*PCA n components* $\in (1, 3, \ \dots\ , 59)$
	$C \in (0.01, 0.1, \ 1, 10)$
	kernel: Radial basis function (RBF)
Random Forest (RF)	*PCA n components* $\in (1, 3, \ \dots\ , 59)$
	n estimators $\in (1, 10, \dots, 200)$
	max depth $\in (1, 3, \dots, 99)$
	max leaf nodes $\in (2, 12, \dots, 92)$
Gaussian Process (GP)	*PCA n components* $\in (1, 3, \ \dots\ , 59)$
	kernel: RBF, WhiteKernel, Matern, DotProduct, ExpSineSquared, ConstantKernel or RationalQuadratic
	Normalize y: True

G298_atom Results

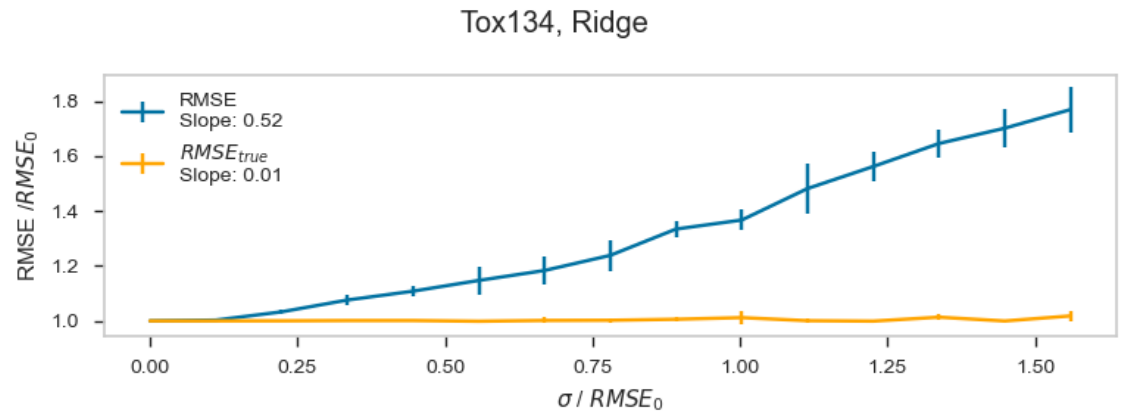


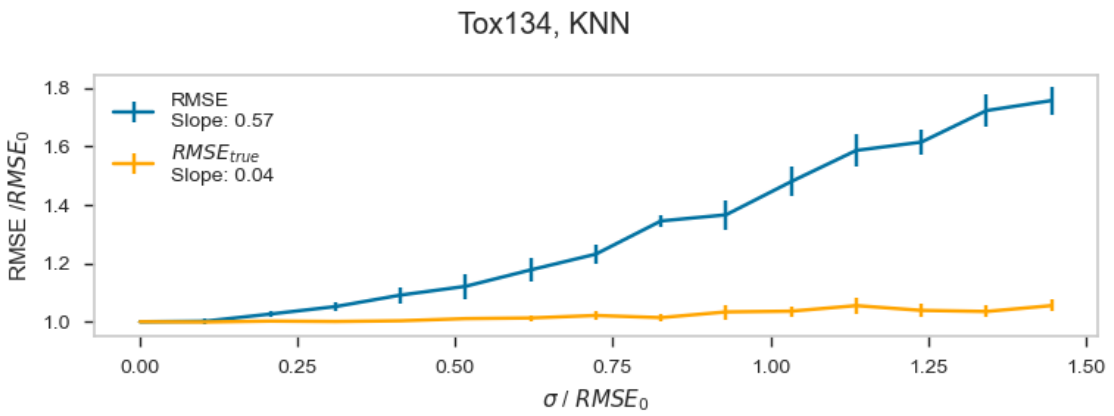


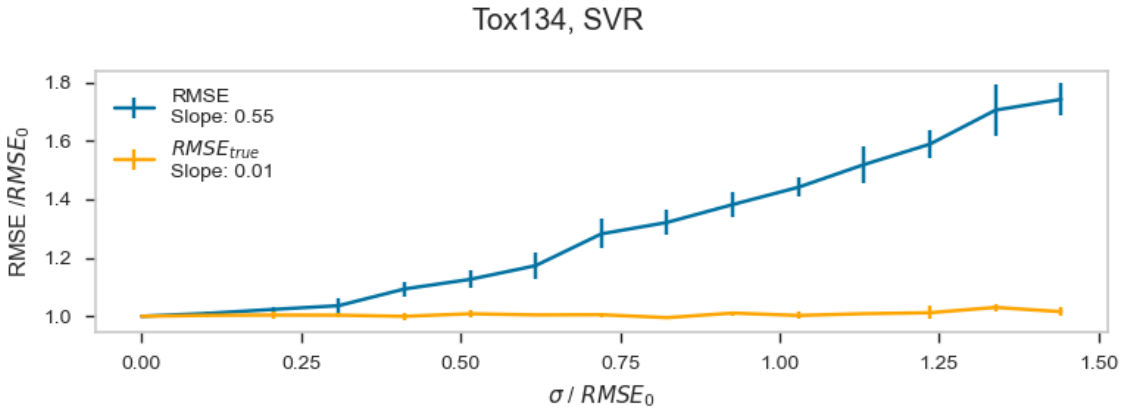


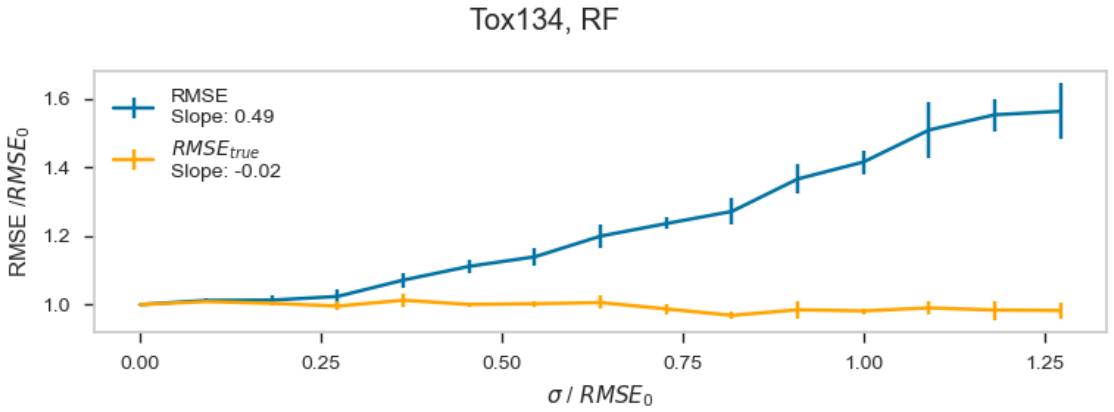


Tox134 Results







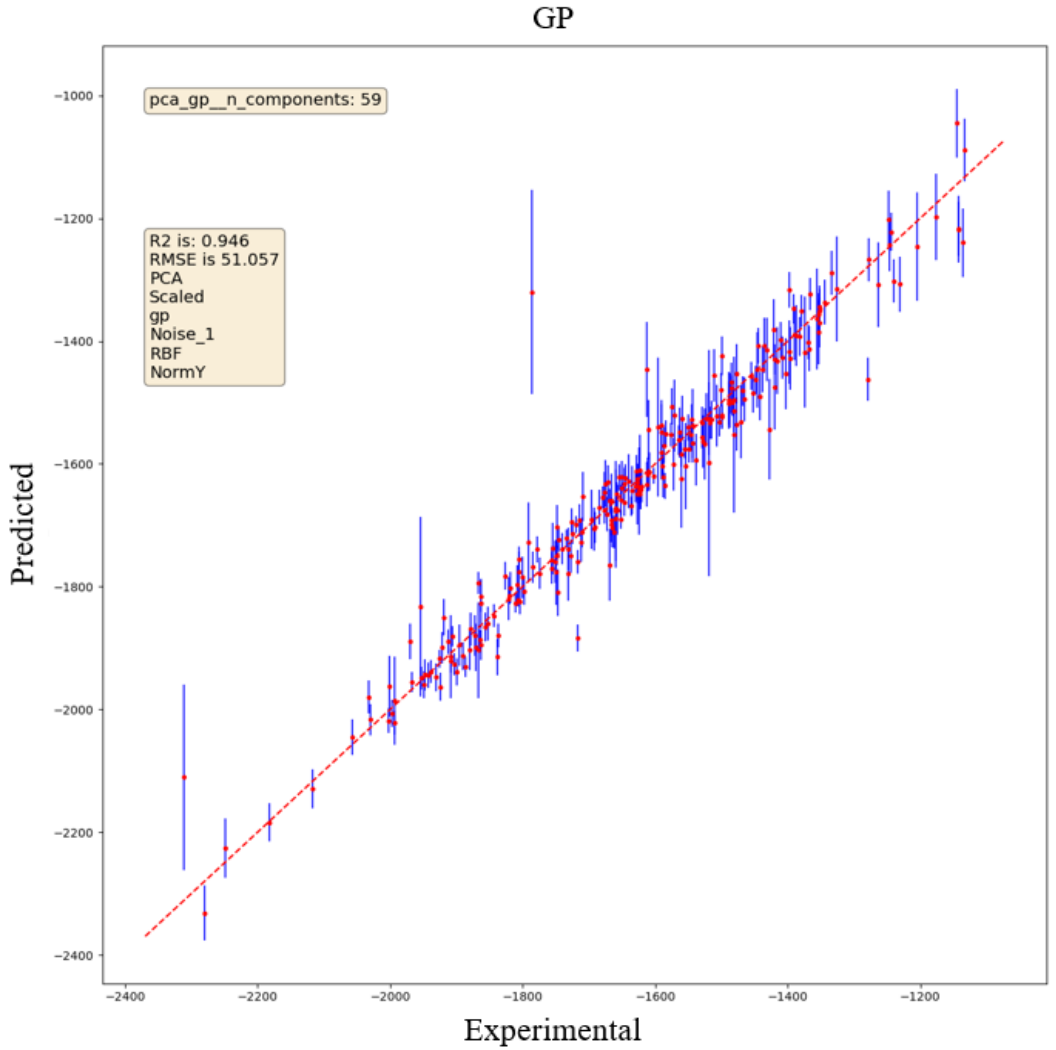


RMSE Slope Ratios

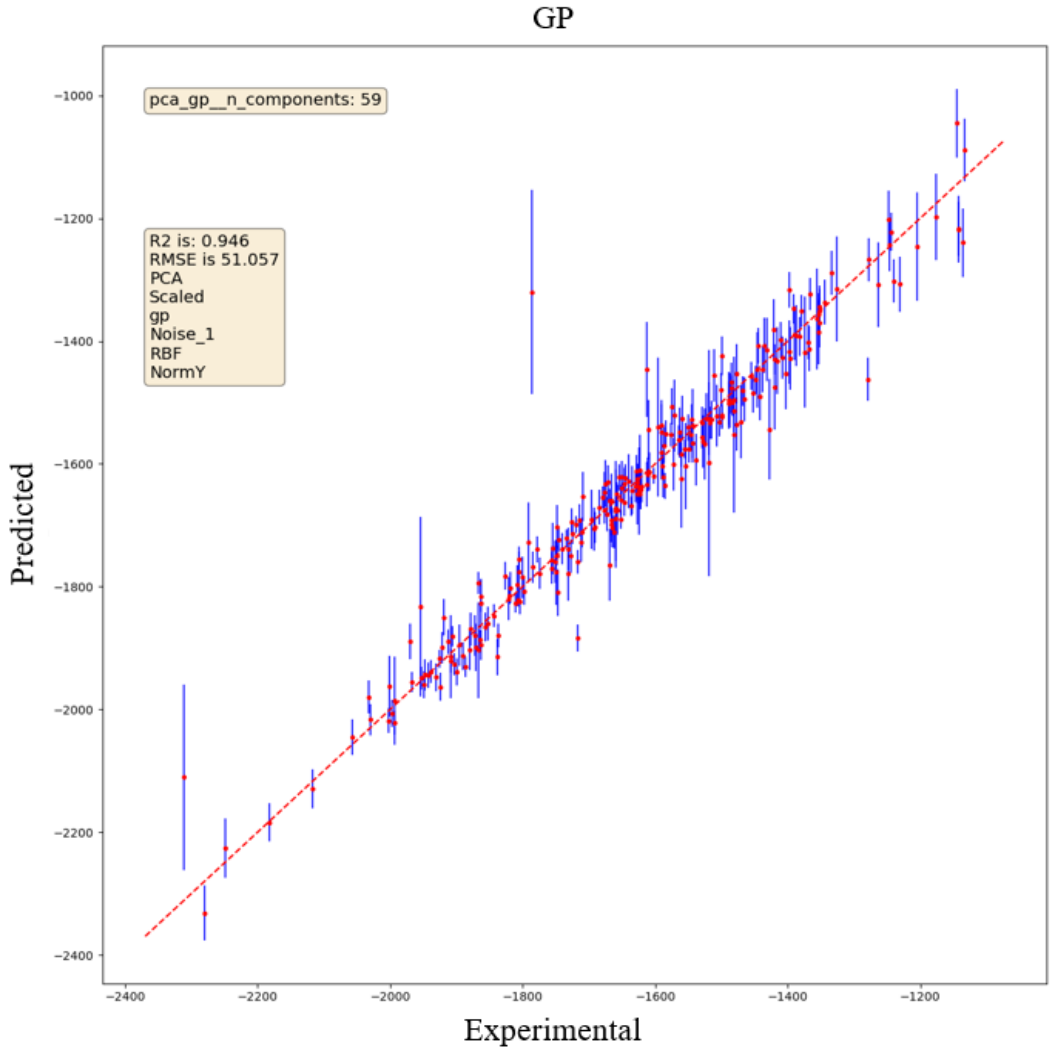
Dataset/Algorithm	Ridge	kNN	SVR	RF	μ ± σ
G_298_atom	5.8	8.8	8.9	8.8	8.1 ± 1.3
Alpha	6.9	8.7	7.3	7.8	7.7 ± 0.67
Lip	19	18	6.9	14	14 ± 4.8
Solv	5.8	3.0	3.3	6.1	4.6 ± 1.4
BACE	13	12	2.9	12	10 ± 4.1
Tox_102	44	10	220	43	79 ± 82
Tox_134	52	14	55	-	40 ± 19
LD50	-	11	6.0	16	11 ± 4.1
μ ± σ	21 ± 18	11 ± 4.1	39 ± 70	15 ± 12	
μ ± σ [a]	10 ± 5.2	10 ± 4.5	5.9 ± 2.1	11 ± 3.5	

[a] With Tox102 and Tox134 ratios omitted.

Gaussian Process (GP) Results



Gaussian Process (GP) Results



$$\hat{Y} = \hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$$

$$\sigma_{\hat{y}} = \sigma_{\hat{y}1}, \sigma_{\hat{y}2}, \ldots, \sigma_{\hat{y}n}$$

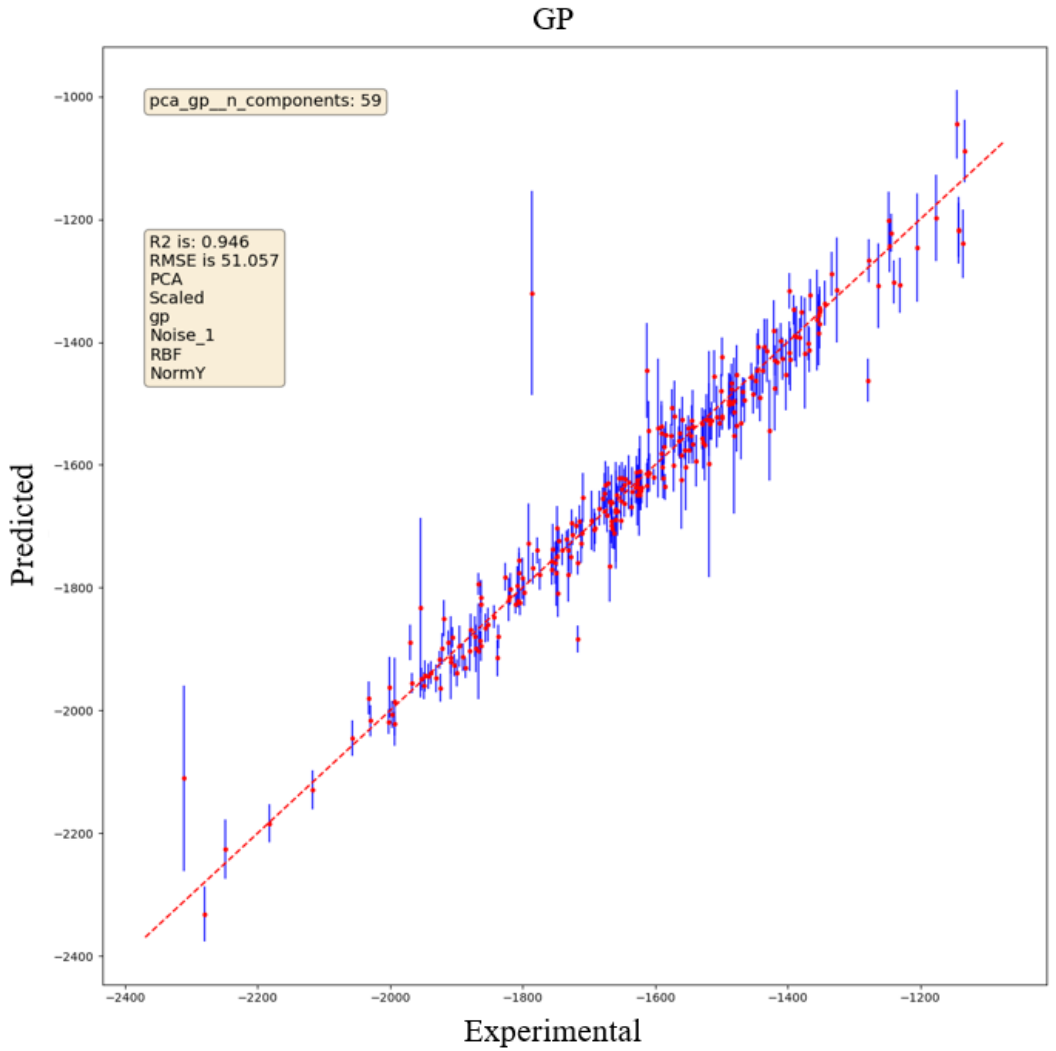
$$Mean\ \sigma_{\hat{y}} = \frac{1}{n}\sum_{i=1}^{n} \sigma_{\hat{y}i}$$

$$\sigma_{\hat{y}}\ 95\%\ CI = \frac{1.960}{\sqrt{n}}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\sigma_{\hat{y}i} - Mean\ \sigma_{\hat{y}}\right)^2\right]$$

$$Y = y_1, y_2, \ldots, y_n$$

$$\sigma_y = \sigma_{y1}, \sigma_{y2}, \ldots, \sigma_{yn}$$

Gaussian Process (GP) Results



$$\hat{Y} = \hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$$

$$\sigma_{\hat{y}} = \sigma_{\hat{y}1}, \sigma_{\hat{y}2}, \ldots, \sigma_{\hat{y}n}$$

$$Mean\ \sigma_{\hat{y}} = \frac{1}{n}\sum_{i=1}^{n} \sigma_{\hat{y}i}$$

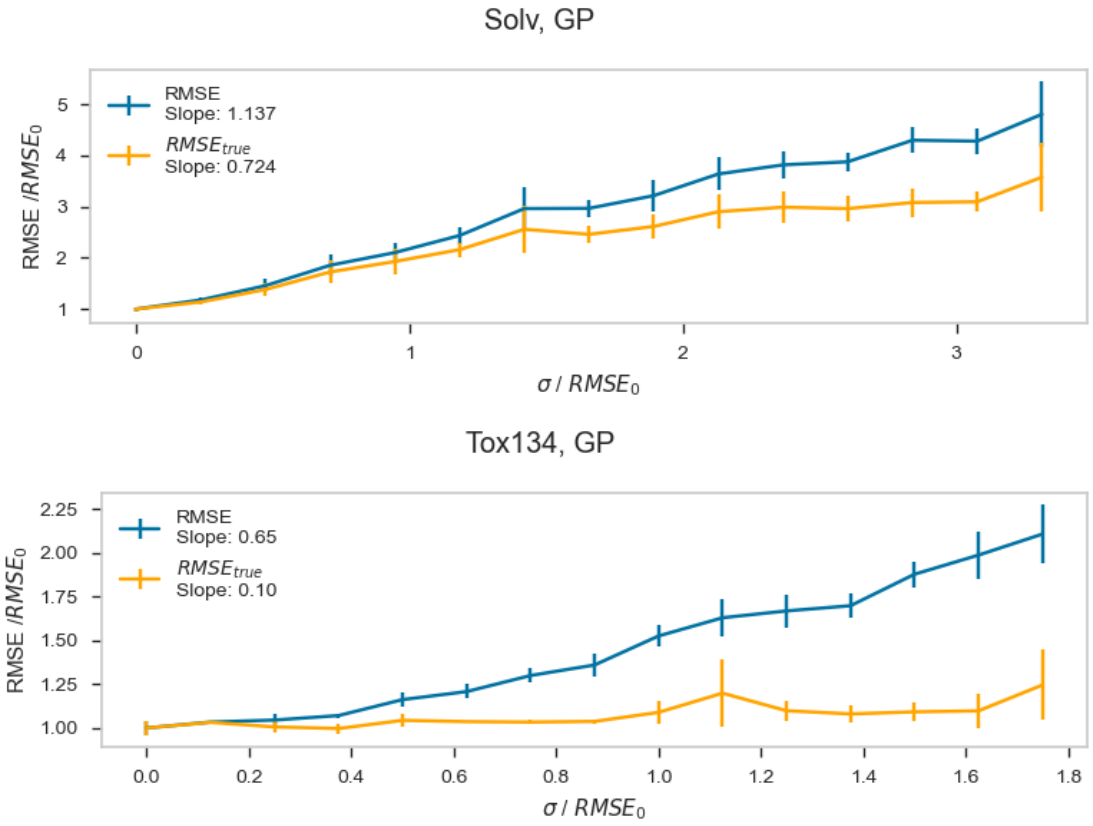
$$\sigma_{\hat{y}}\ 95\%\ CI = \frac{1.960}{\sqrt{n}}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\sigma_{\hat{y}i} - Mean\ \sigma_{\hat{y}}\right)^2\right]$$

$$Y = y_1, y_2, \ldots, y_n$$

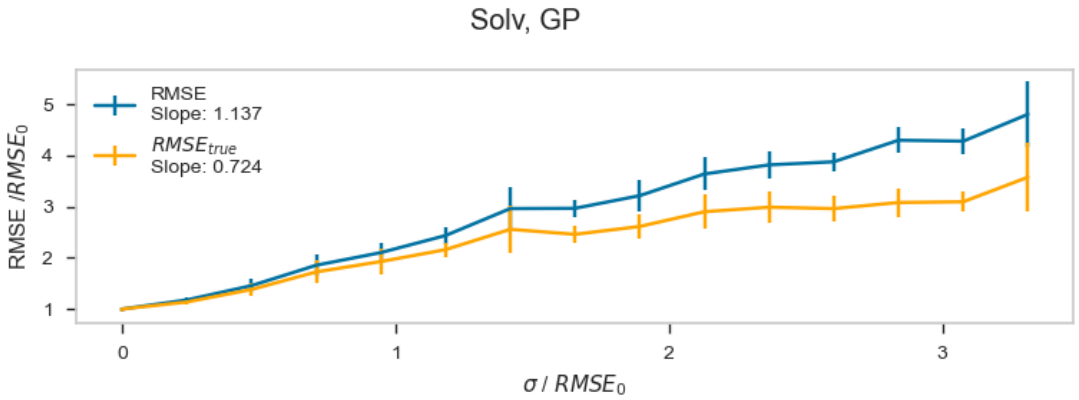
$$\sigma_y = \sigma_{y1}, \sigma_{y2}, \ldots, \sigma_{yn}$$

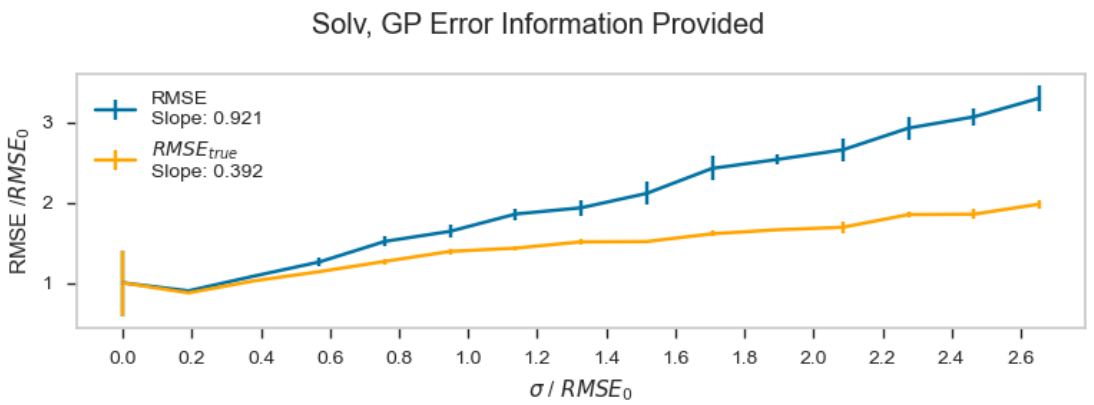
Information about experimental uncertainty

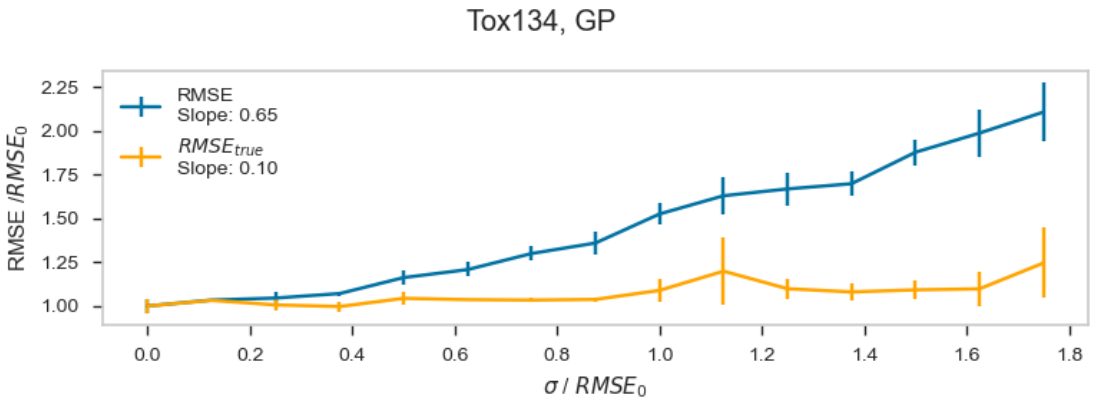
Gaussian Process (GP) Results

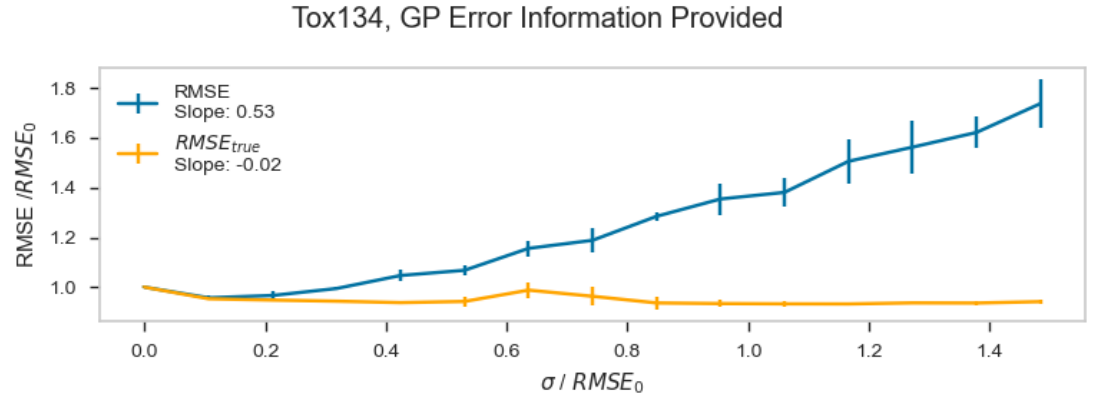


Gaussian Process (GP) Results









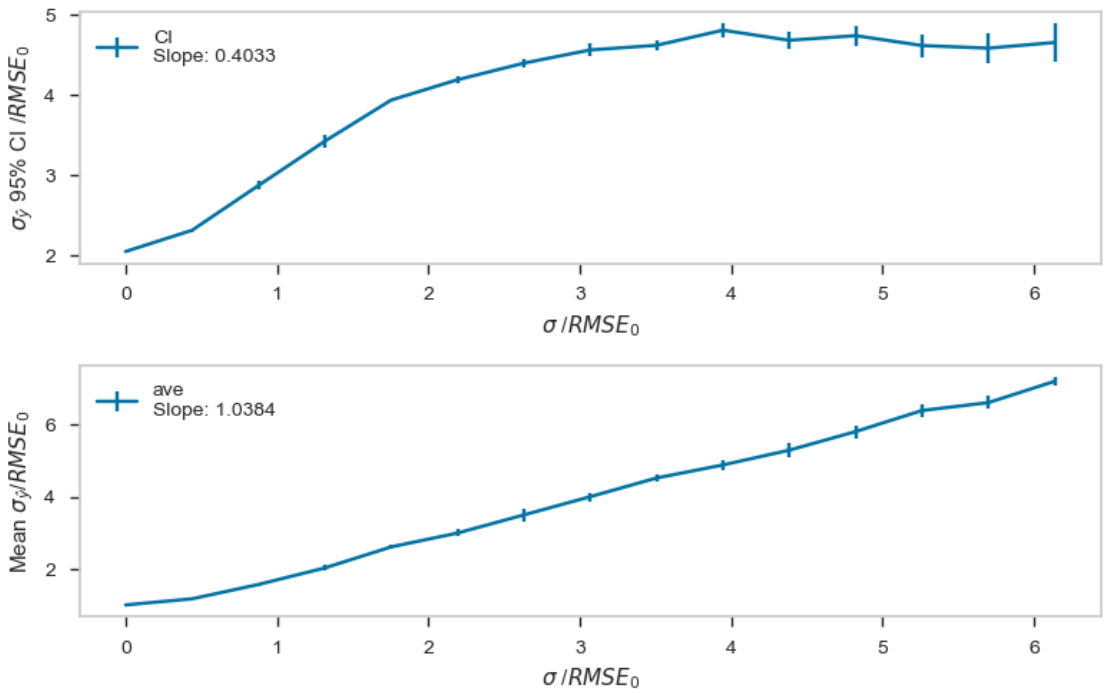
GP Slope ratios

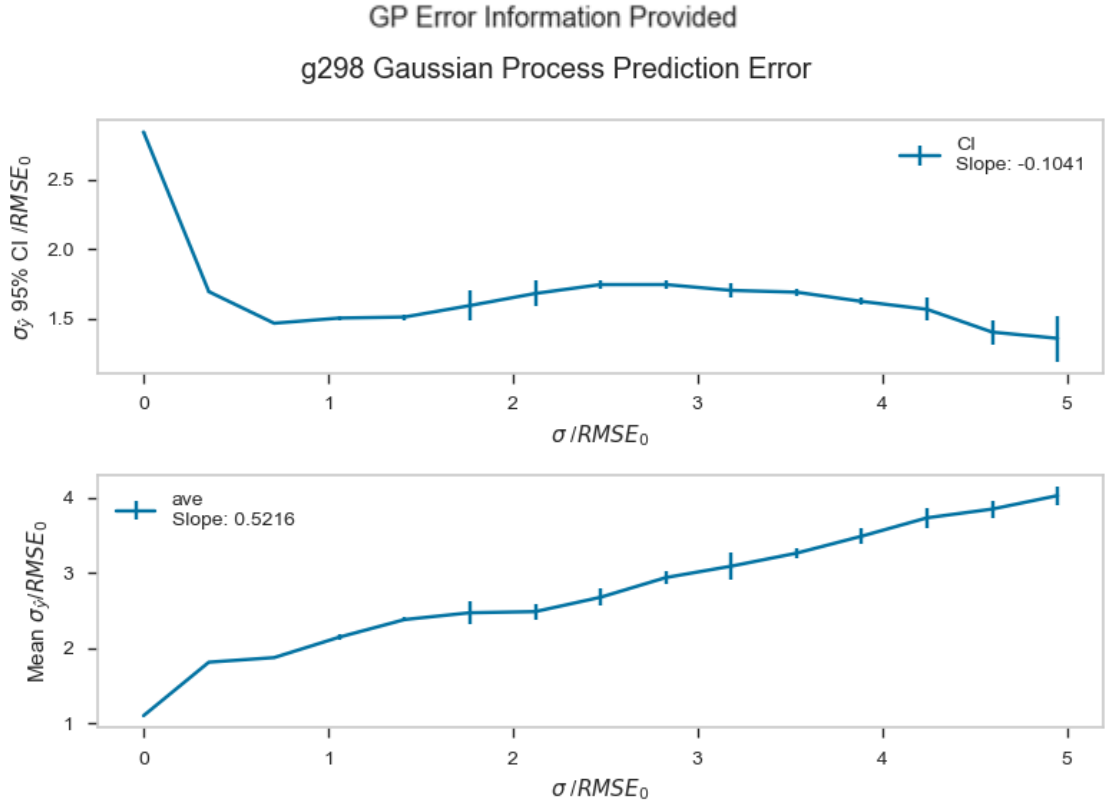
Dataset	*No* σ_y	*With* σ_y
G_298_atom	1.9	2.0
Alpha	1.8	9.4[a]
Solv	1.6	2.5[a]
BACE	3.8	7.8[a]
Tox_102	2.8	-[b]
Tox_134	7.0	-[b]
LD50	5.4	6.0
$\mu \pm \sigma$	3.5 ± 1.9	5.5 ± 2.9

[a]Slopes m and m_{true} were calculated excluding the first two points due to a discontinuity in the line.

[b]The slope m_{true} was negative for these plots, so the slope ratio was not calculated.

Gaussian Process (GP) Results



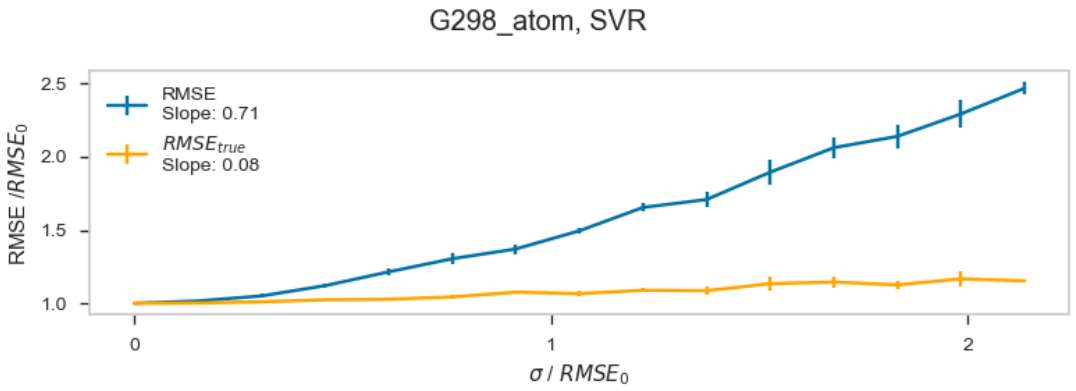


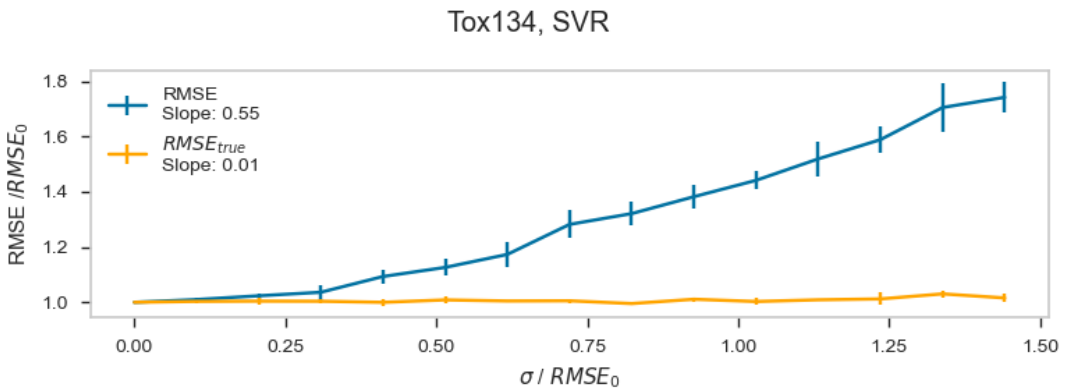
GP Prediction Uncertainties

Dataset	*No* σ_y *Mean* $\sigma_{\hat{y}}$	*No* σ_y $\sigma_{\hat{y}}$ *95% CI*	*With* σ_y *Mean* $\sigma_{\hat{y}}$	*With* σ_y $\sigma_{\hat{y}}$ *95% CI*
G_298_atom	1.0	0.40	0.52	-0.10
Alpha	1.1	0.16	0.44[a]	0.32[a]
Solv	0.94	-0.19	0.10	0.10
BACE	0.25	0.38	-0.12	-0.35
Tox_102	0.32	0.028	-0.96	-0.48
Tox_134	0.49	0.53	-0.66	-0.17
LD50	0.66	-0.39	-0.60	0.14

- "It follows that the model's prediction of the *external test set* will have uncertainty equal to or greater than that contained within the *training set"*
- Because these slopes are < 1, prediction uncertainty is actually *less* than the added error!

Conclusions

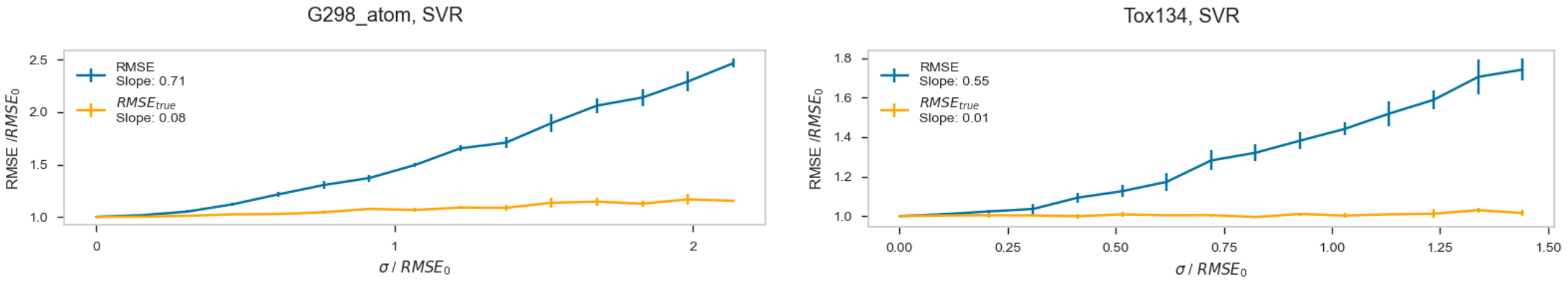




Methods

- Gaussian error was added to *8 representative QSAR datasets* and modeled using *5 algorithms*
  - The use of Gaussian distributed error represents an *ideal* but *realistic* simulation of real-world modeling

Conclusions



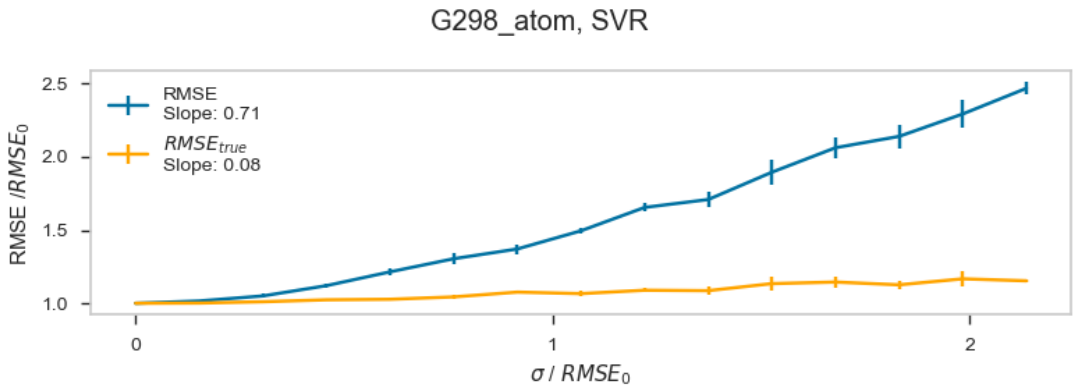
Methods

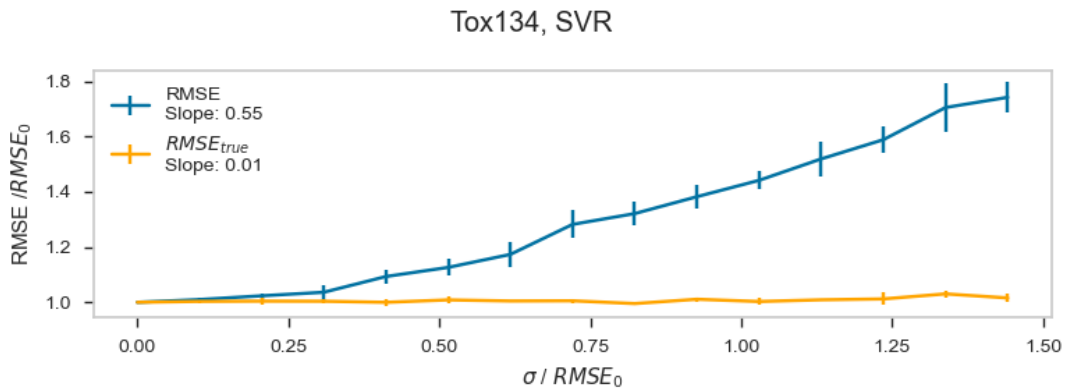
- Gaussian error was added to *8 representative QSAR datasets* and modeled using *5 algorithms*
  - The use of Gaussian distributed error represents an *ideal* but *realistic* simulation of real-world modeling

Results

- For each dataset and algorithm, the *true test set* was always predicted more accurately than the *error laden test set*
- The difference between *RMSE* and $RMSE_{true}$ depends on algorithm, dataset, and the level of added error
- When using Gaussian Process
  - Increasing the simulated error increases the prediction uncertainty
  - Providing information about error to the algorithm mitigates these trends
  - **Prediction uncertainty is often *less* than the amount of <u>added</u> error!**

Conclusions

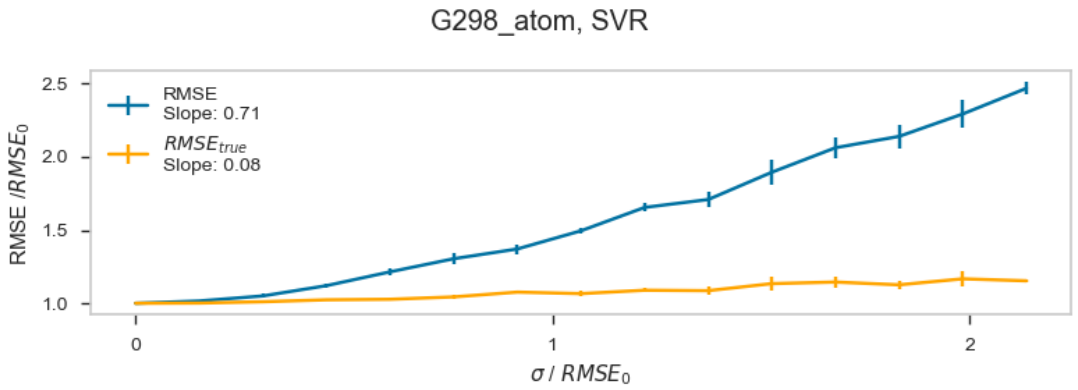


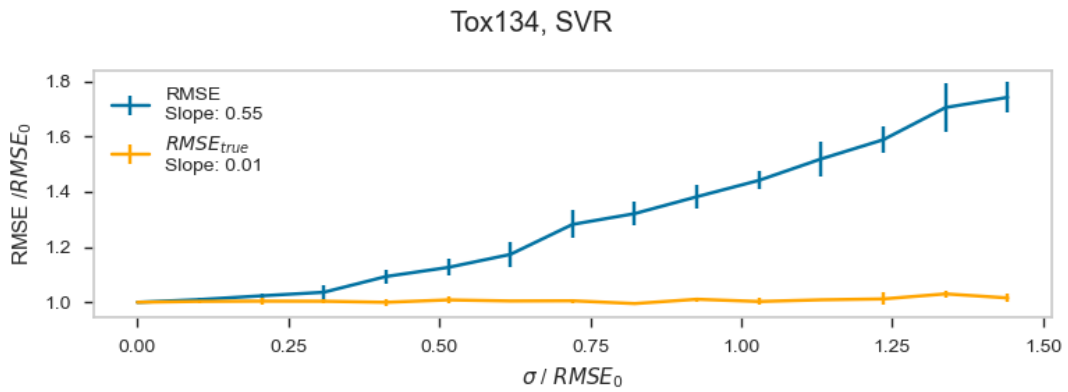


Implications

- QSAR models *can* make predictions which are more accurate than training data
- Evaluation of QSAR models on error laden test sets can give flawed interpretations of performance
  - A model may be making good predictions, but this will be obscured by test set error

Conclusions





Implications

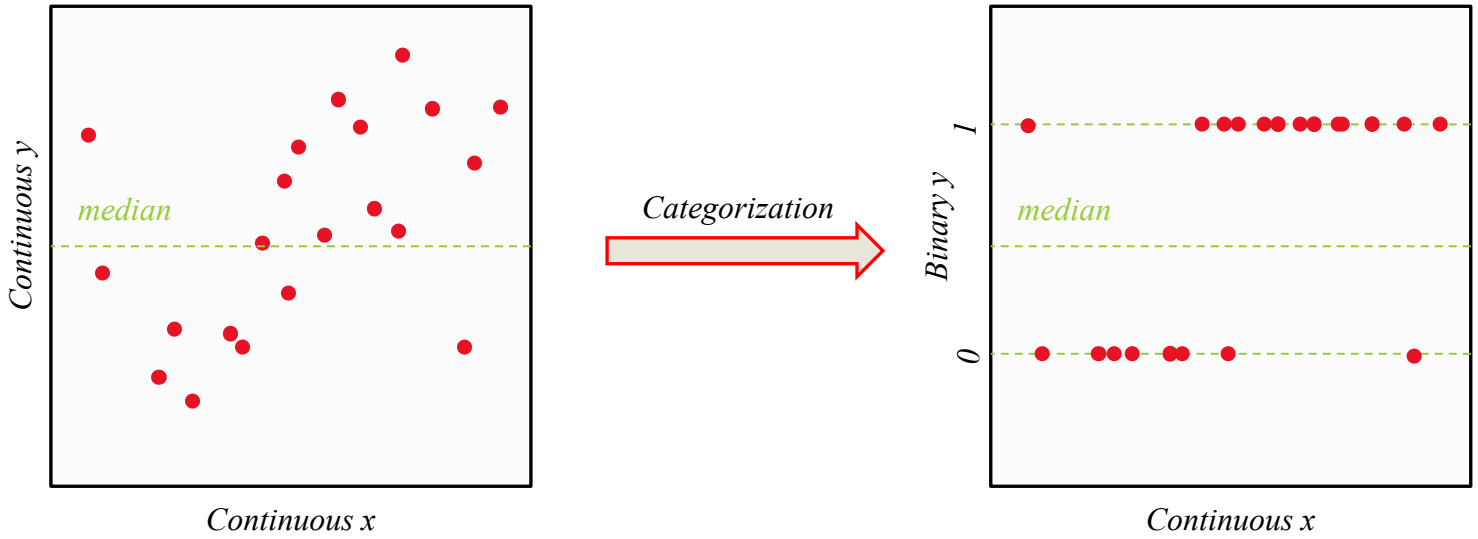
- QSAR models *can* make predictions which are more accurate than training data
- Evaluation of QSAR models on error laden test sets can give flawed interpretations of performance
  - A model may be making good predictions, but this will be obscured by test set error
- Different models respond differently to error
  - $RMSE/RMSE_{true}$ is model dependent
  - $RMSE$ is observed
  - $RMSE_{true}$ is unknown
- Determining relative performance can be tenuous and potentially misleading

Future Work

- Evaluation of new algorithms and new models will be similarly limited by knowledge of the uncertainty in validation and test sets
- New methods of inferring uncertainty in datasets and new evaluation methodologies which utilize knowledge of uncertainty are needed to give more reliable comparisons of QSAR models
- Our group will focus on sources of error prominent in toxicological modeling, particularly systematic error

Categorizing Continuous Data in QSAR

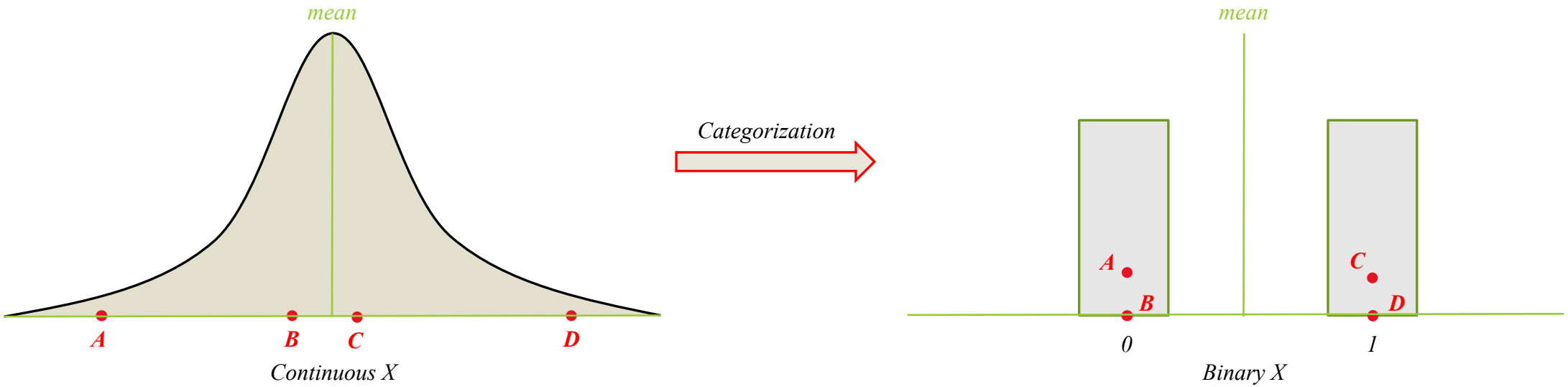
Splitting Data



Cheminformaticians often split (categorize) *continuous data* into *categorical data*

But this leads to a *loss of information*, *loss of effect size*, and *loss of statistical significance* between variables

Loss of Information



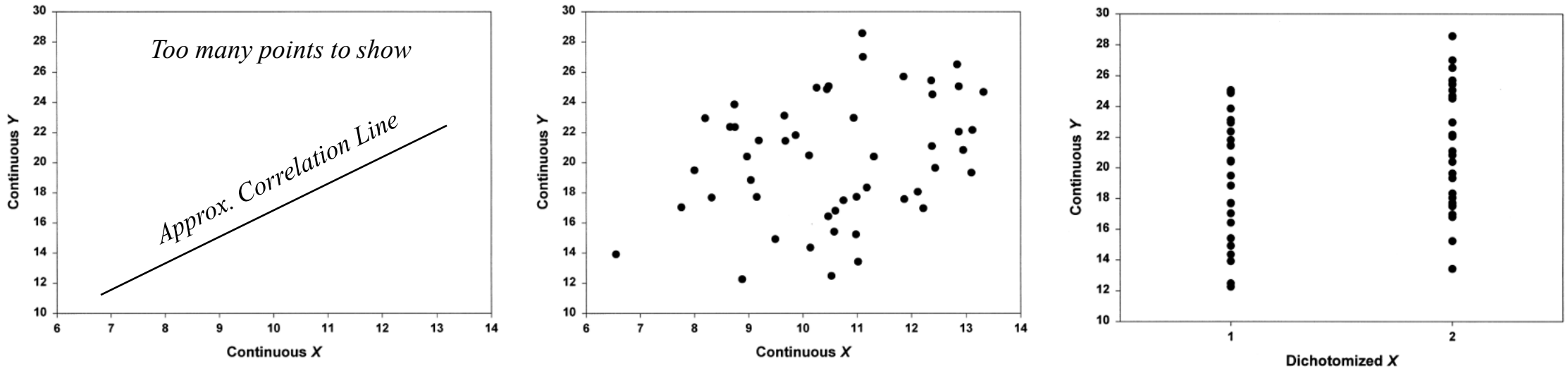
Scenario:

- *C* is closer to *B* than to *D*

Result:

- Loss of individual differences between observations
- *C* and *D* are judged to be more similar than *C* and *B*

Loss of Effect Size and Statistical Significance



Population

- $n = > 1x10^6$
- $\rho_{xy} = 0.40$

Sample

- $n = 50$
- $r_{xy} = 0.30$
- *95% CI* = [0.02, 0.53]
- *Null Hypothesis*: $\rho_{xy} = 0.0$
- $t(48) = 2.19$, $p = 0.03$

Same Sample

- $n = 50$
- $r_{xy} = 0.21$
- *95% CI* = [-0.07, 0.46]
- *Null Hypothesis*: $\mu_1 = \mu_2$
- $t(48) = 1.47$, $p = 0.15$

MacCallum et al. *Psychological Methods*, **2002**, *7*, 19.

Splitting Up: It's a Bad Idea..

STATISTICS IN MEDICINE
Statist. Med. 2006; **25**:127–141
Published online 11 October 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2331

Dichotomizing continuous predictors in multiple regression: a bad idea

Patrick Royston[1,*,†], Douglas G. Altman[2] and Willi Sauerbrei[3]

Psychological Methods
2002, Vol. 7, No. 1, 19–40

1082-989X/02/$5.00 DOI: 10.1037//1082-989X.7.1.19

On the Practice of Dichotomization of Quantitative Variables

Robert C. MacCallum, Shaobo Zhang, Kristopher J. Preacher, and Derek D. Rucker
Ohio State University

described, and justifications that are offered for such usage are examined. The authors present the case that dichotomization is rarely defensible and often will yield misleading results.

JULIE R. IRWIN and GARY H. McCLELLAND*

Marketing researchers frequently split (dichotomize) continuous predictor variables into two groups, as with a median split, before performing data analysis. The practice is prevalent, but its effects are not well understood. In this article, the authors present historical results on the effects of dichotomization of normal predictor variables rederived in a regression context that may be more relevant to marketing researchers. The authors then present new results on the effect of dichotomizing continuous predictor variables with various nonnormal distributions and examine the effects of dichotomization on model specification and fit in multiple regression. The authors conclude that dichotomization has only negative consequences and should be avoided.

Negative Consequences of Dichotomizing Continuous Predictor Variables

Dichotomizing Continuous Outcome Variables: Dependence of the Magnitude of Association and Statistical Power on the Cutpoint

David R. Ragland

Dichotomizing a continuous outcome variable casts that variable in traditional epidemiologic terms (that is, disease, no disease). One consequence is overall reduced statistical power. A more fundamental concern is that the magnitude

Finding What Is Not There through the Unfortunate Binning of Results: The Mendel Effect

Howard Wainer, Marc Gessaroli, and Monica Verdi
National Board of Medical Examiners

Splitting a Predictor at the Upper Quarter or Third and the Lower Quarter or Third

Andrew GELMAN and David K. PARK

uniformly or normally distributed. By discretizing x into three categories, we claw back about half the efficiency lost by the commonly used strategy of dichotomizing the predictor.

Project Plan

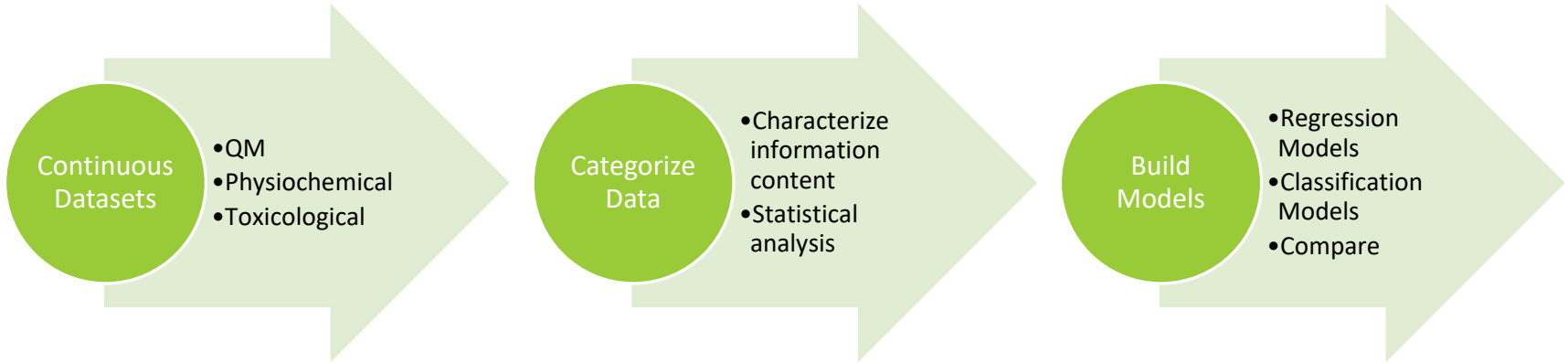
Hypothesis

Categorization of continuous data is bad statistical practice
and distorts the relationship between variables.

Will this fundamental principle result in less predictive machine learning models?
How does categorization affect the quality of predictions in QSAR?

Approach

Using continuous datasets, make predictions *before* (Regression) and *after* (Classification) categorization.



Datasets

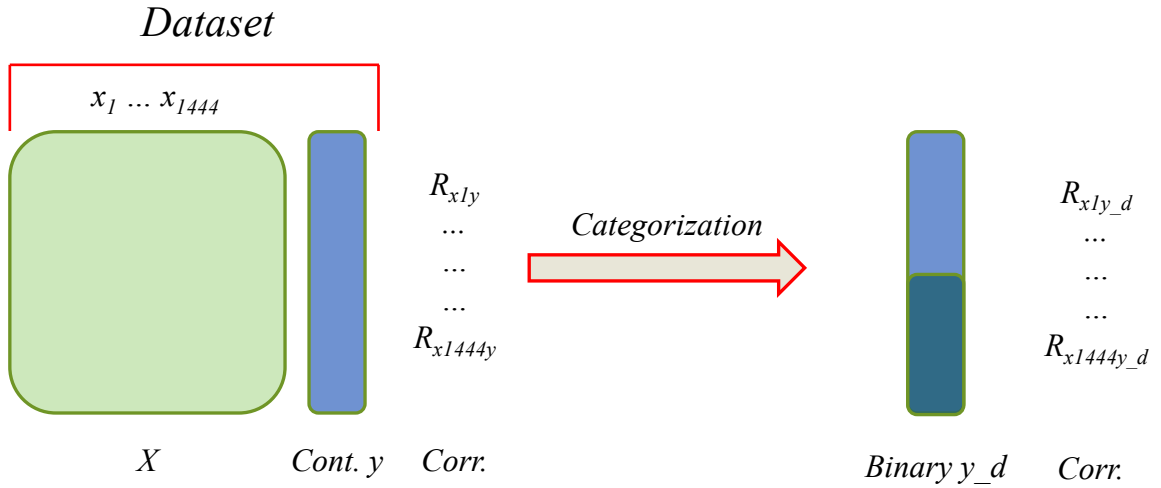
Dataset	Category	Number of Molecules[a]	Endpoint	Range
G298_atom	Quantum Mechanical	131,082	ΔG^{o}_{at} (kcal mol^{-1})	-2,417 – -288
Solv	Physiochemical	642	ΔG^{o}_{hyd} (kcal mol^{-1})	-25.5 – 3.4
Tox_102[b]	Toxicological *in vitro*	971	$logAC_{50}$	-2.1 – 4.7
Tox_134[c]	Toxicological *in vitro*	1,347	$logAC_{50}$	-4.0 – 2.8

[a] Original size of the dataset. If datasets have more than 1,000 molecules, they were randomly sampled down to a size of 1,000 before modeling.

[b] Includes data exclusively from the ATG-PPre-cis assay

[c] Inclues data exclusively from the ATG-PPARg-trans assay

Loss of Correlation Effect Size

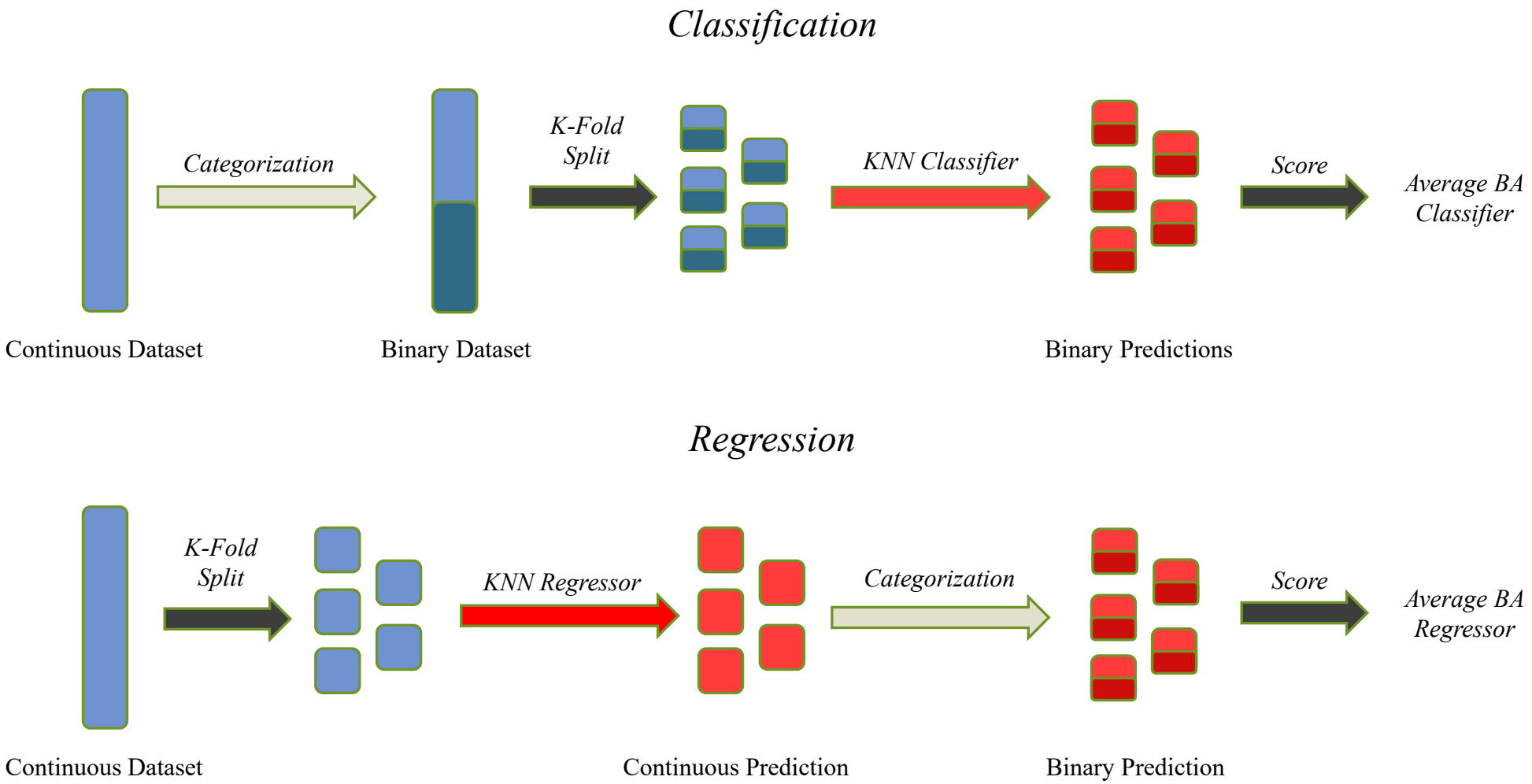


Statistics for the Difference in Correlation Vectors $R_{xy\_d} - R_{xy}$

	mean	std	min	25th perc	50th perc	75th perc	max	Num. of R with p < 0.05
g298atom	-0.020	0.079	-0.22	-0.061	-0.0078	0.017	0.25	-9
Solv	-0.021	0.060	-0.27	-0.054	-0.019	0.015	0.26	-10
Tox102[a]	-0.015	0.036	-0.010	-0.042	-0.014	0.014	0.068	-33
Tox134[a]	-0.021	0.031	-0.11	-0.042	-0.019	0.0011	-0.089	-53

[a] Log10 scale

Comparing Classification and Regression



Modeling Methods

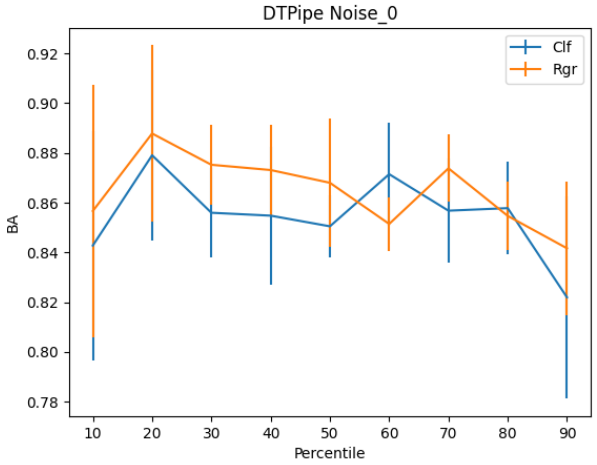
Compare 3 Modeling Approaches

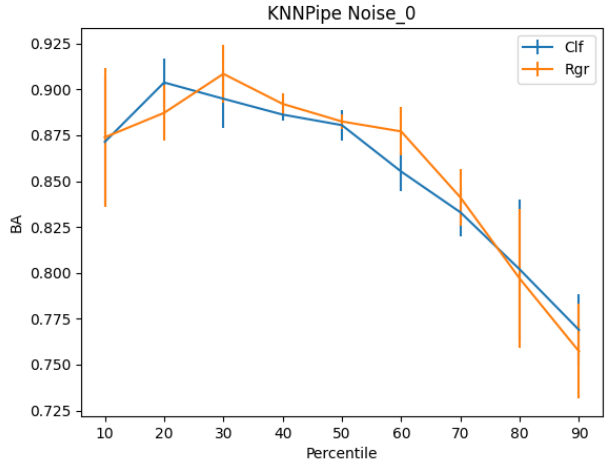
- *Standard*
  - Algorithm

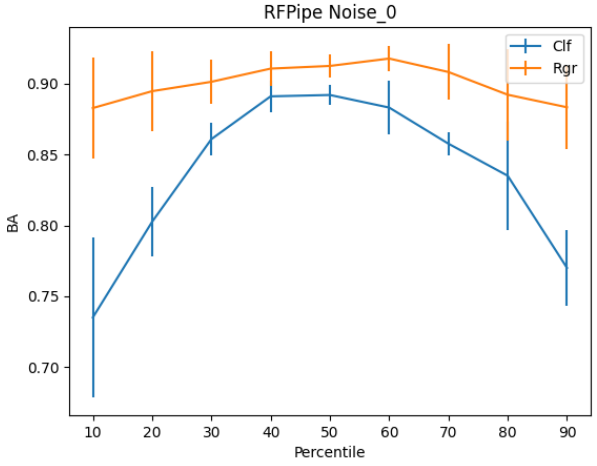
- *Pipe*
  - Standard Scaler
  - PCA *n* = *x*
  - Algorithm

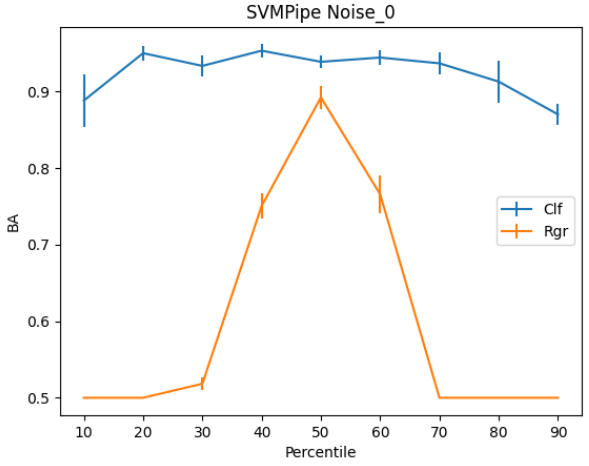
- *Opt*
  - Standard Scaler
  - GridSearchCV
    - DT
      - PCA *n*
      - Max Depth
      - Min Sample Split
      - Min Sample Leaf
    - KNN
      - PCA *n*
      - KNN *n*
    - RF
      - PCA *n*
      - N Estimators
      - Max Depth
      - Min Sample Split
      - Min Sample Leaf
    - SVM
      - PCA *n*
      - Kernel
      - C
  - PCA *n* = *optimized*
  - Algorithm

Results – *G298Atom*





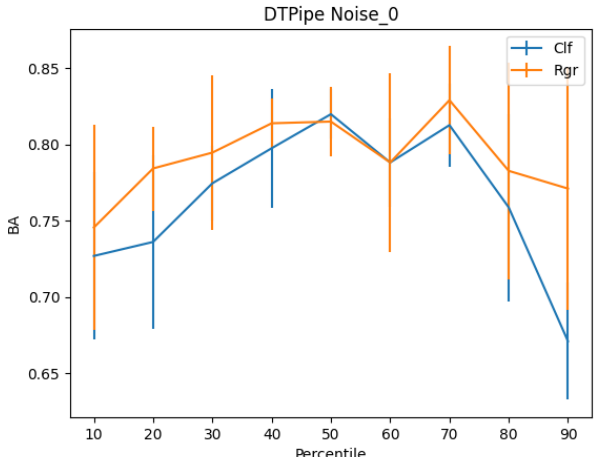


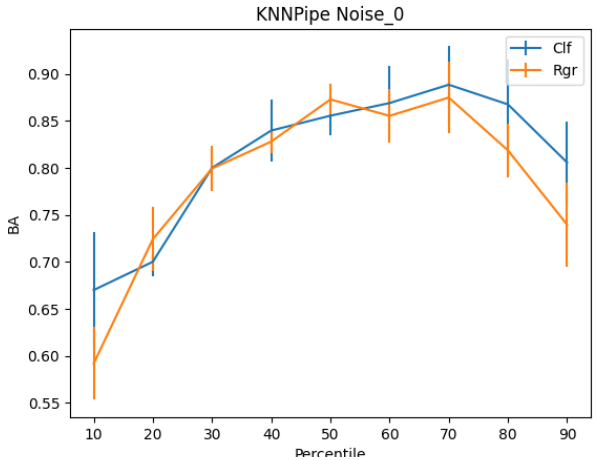


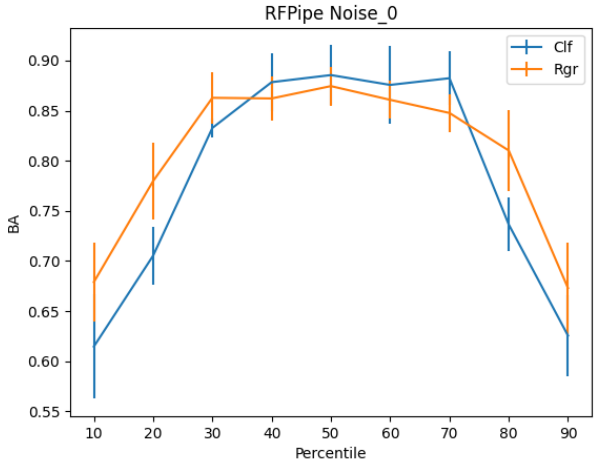
Results Overview

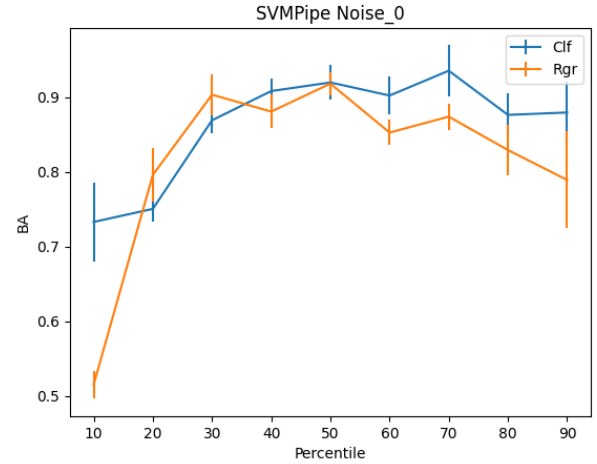
- *y*-axis is Balanced Accuracy
  - Blue lines are classifier scores
  - Orange lines are regressor scores
  - μ±σ of five 5-fold splits (stratified)
- *x*-axis is the percentile of the target variable at which the data was binarized
- *Pipe* modeling method
  - PCA *n* = 100

Results - *Solv*





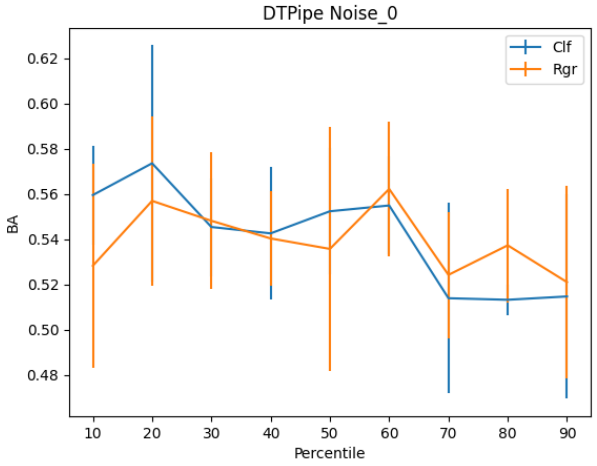


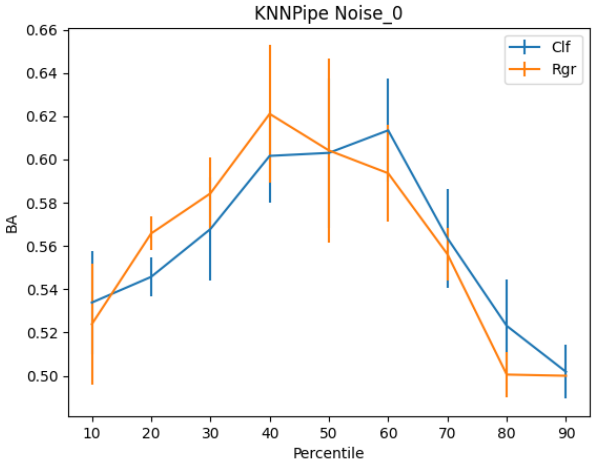


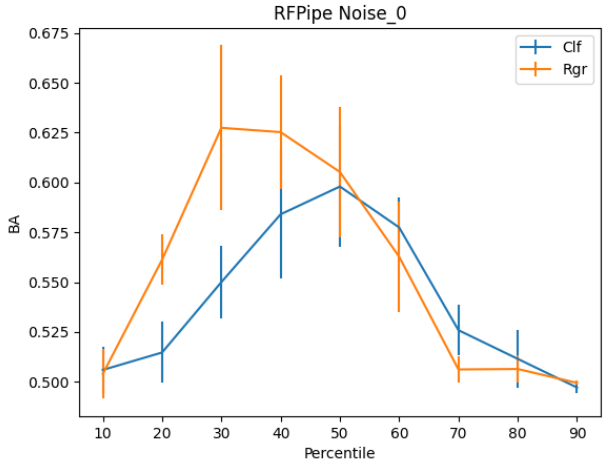
Results Overview

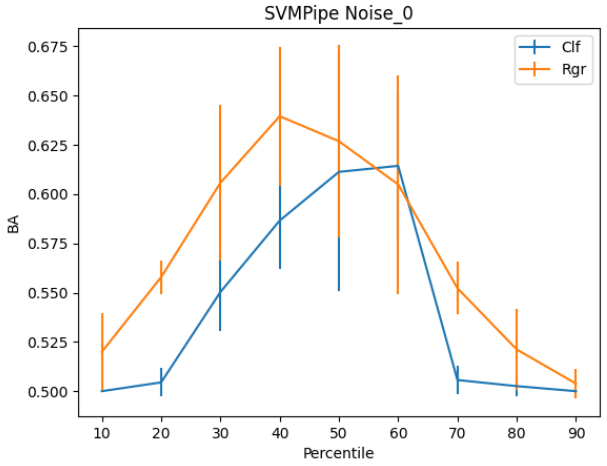
- *y*-axis is Balanced Accuracy
  - Blue lines are classifier scores
  - Orange lines are regressor scores
  - μ±σ of five 5-fold splits (stratified)
- *x*-axis is the percentile of the target variable at which the data was binarized
- *Pipe* modeling method
  - PCA *n* = 100

Results – *Tox102*





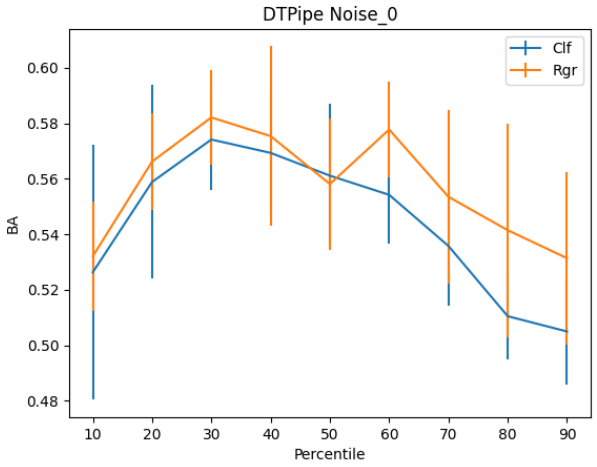


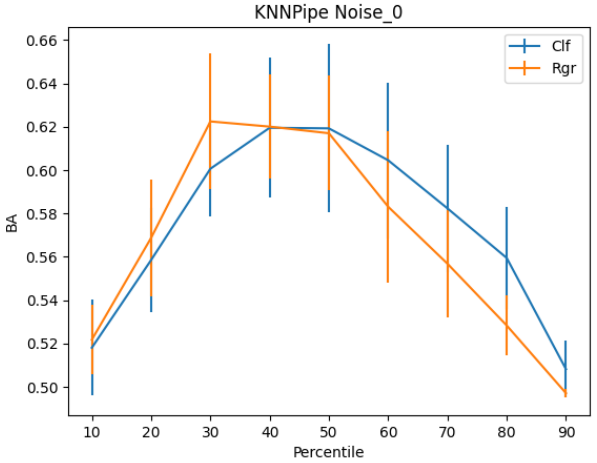


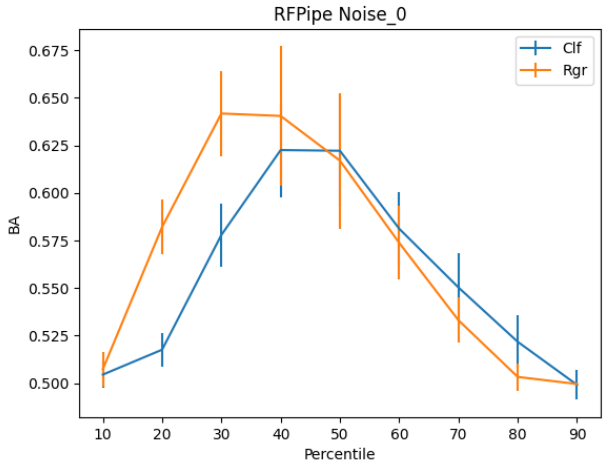
Results Overview

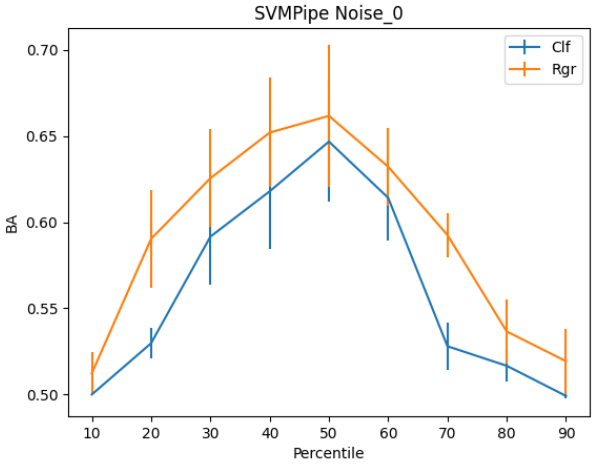
- *y*-axis is Balanced Accuracy
  - Blue lines are classifier scores
  - Orange lines are regressor scores
  - μ±σ of five 5-fold splits (stratified)
- *x*-axis is the percentile of the target variable at which the data was binarized
- *Pipe* modeling method
  - PCA $n = 100$

Results *Tox134*









Results Overview

- *y*-axis is *Balanced Accuracy*
  - Blue lines are classifier scores
  - Orange lines are regressor scores
  - μ±σ of five 5-fold splits (stratified)
- *x*-axis is the percentile of the target variable at which the data was binarized
- *Pipe* modeling method
  - PCA *n* = 100

Performance Metrics

Hypothesis

Balanced Accuracy (*BA*) is not a perfect or comprehensive metric of performance.

How will relative performance between the classification and regression methods compare if other metrics, including probabilistic metrics, are included?

Approach

Aggregate several metrics and compare performance.

Classification Performance Metrics

Confusion Matrix

	Actual Negative	Actual Positive
Predicted Negative	**TN**	*FN*
Predicted Positive	*FP*	**TP**

$$Recall = Sensitivity = True\ Positive\ Rate = \frac{TP}{TP + FN}$$

$$Specificity = True\ Negative\ Rate\ = \frac{TN}{TN + FP}$$

$$Precision = \frac{TP}{TP + FP}$$

- Good for when *false negatives* are important
- How many *actual* negatives were predicted correctly
- Good for when *false positives* are important

Classification Performance Metrics

Confusion Matrix

	Actual Negative	Actual Positive
Predicted Negative	**TN**	*FN*
Predicted Positive	*FP*	**TP**

$$Balanced\ Accuracy = \frac{Recall + Specificity}{2}$$

- Arithmetic mean of *Recall* and *Specificity*, more holistic than both

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

- Harmonic mean of *Precision* and *Recall*, more holistic than both

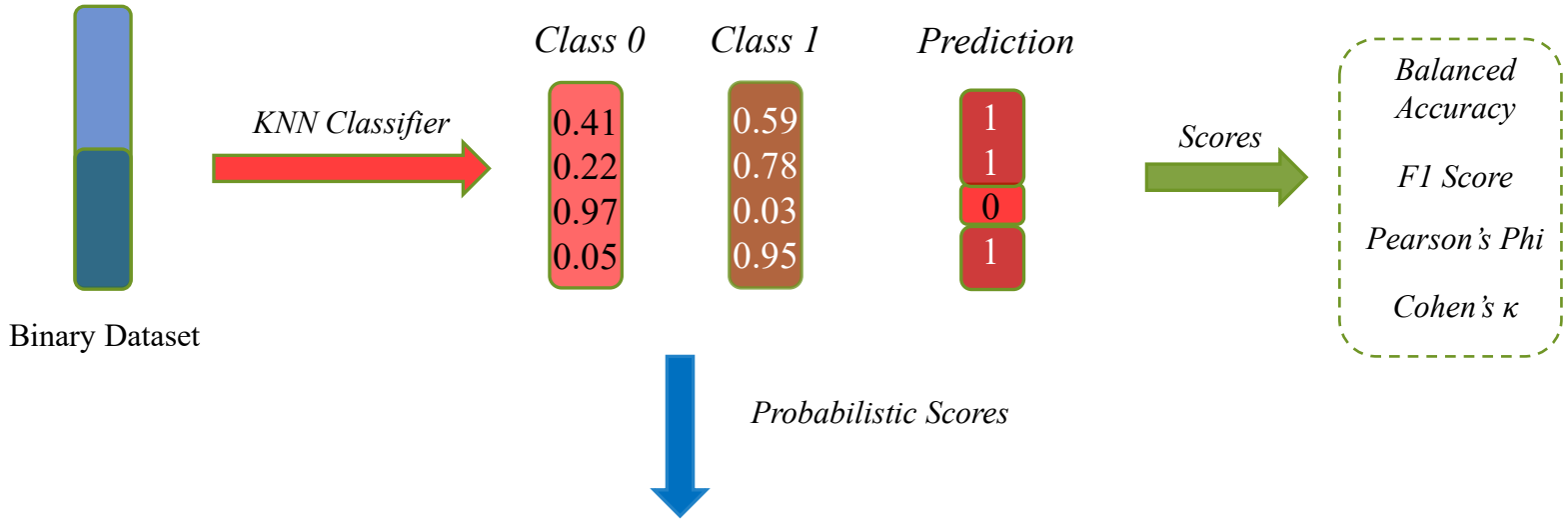
$$Pearson's\ \varphi = \sqrt{\frac{\chi^2}{n}} = \frac{TN * TP - FN * FP}{\sqrt{(TN + FN)(FP + TP)(TN + FP)(FN + TP)}}$$

- Degree of association between predicted and experimental outcomes

$$Cohen's\ \kappa = \frac{p_o - p_e}{1 - p_e}$$

- Agreement between predicted and experimental outcomes, also corrects for probability of random agreement

Probabilistic Performance Metrics



ROC-AUC

- Performance function
- Adjusts probability cutoff for determining class predictions
- Calculate area under the resulting curve
- Higher is better

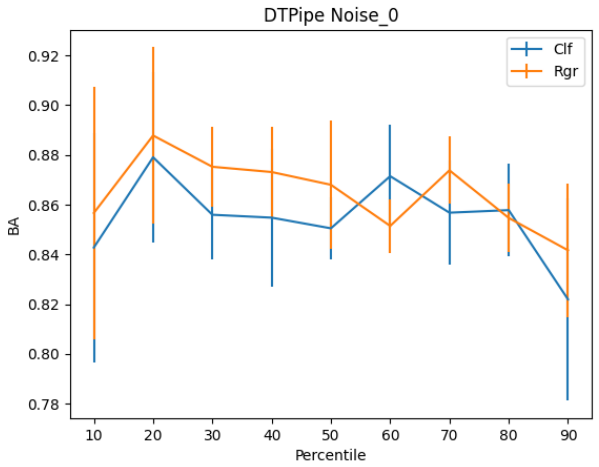
Log Loss

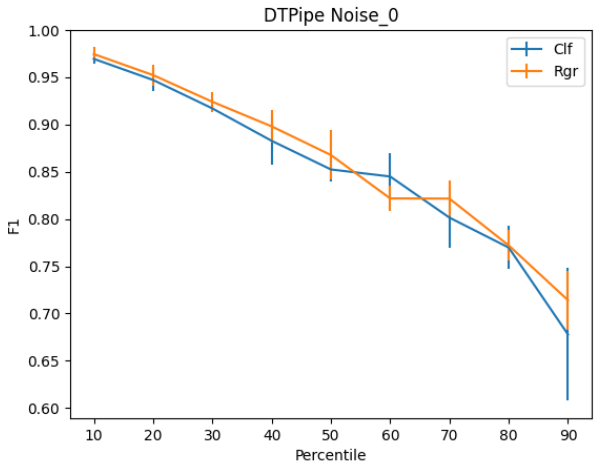
- Cost function
- Logarithm of a loss function incorporating probabilities
- Penalizes for very confident wrong predictions
- Lower is better

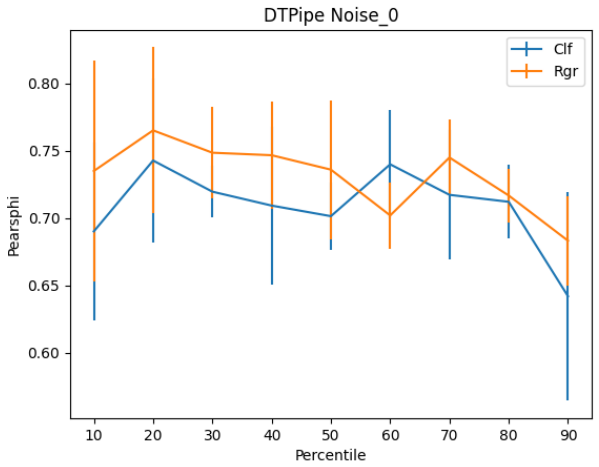
Brier Score

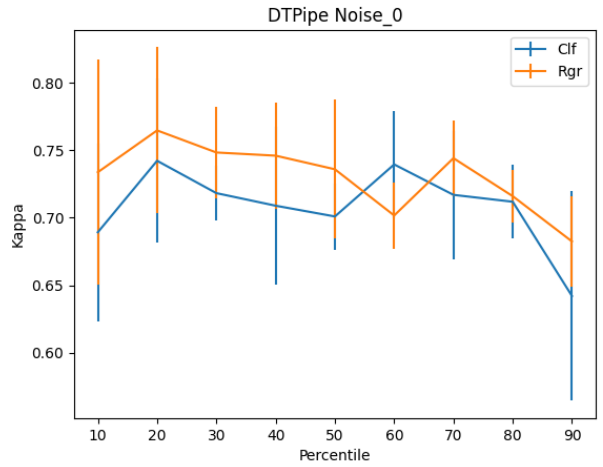
- Cost function
- Mean squared difference between the probability of a prediction and its label
- Also penalizes for very confident wrong predictions
- Lower is better

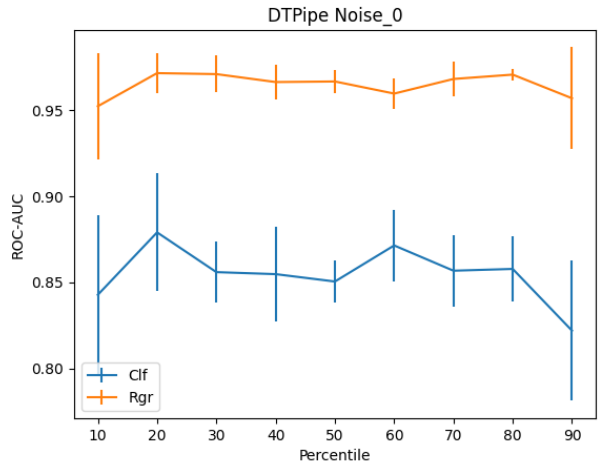
Multiple Metrics – *G298Atom + DT*

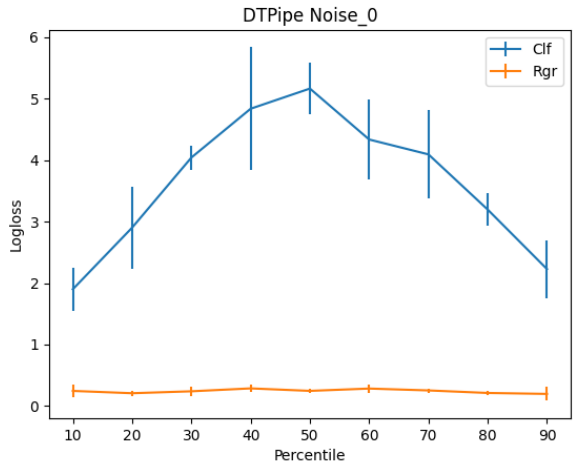


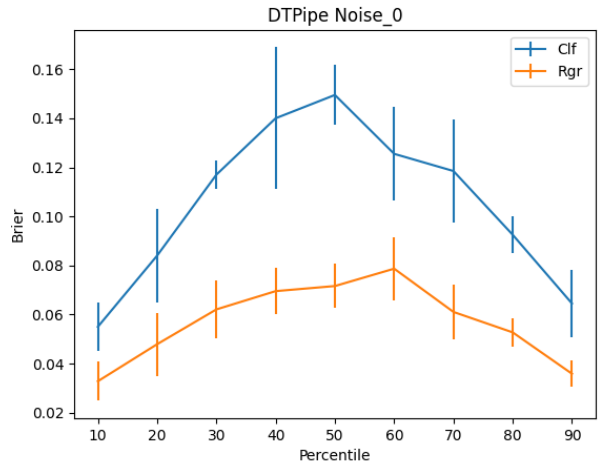




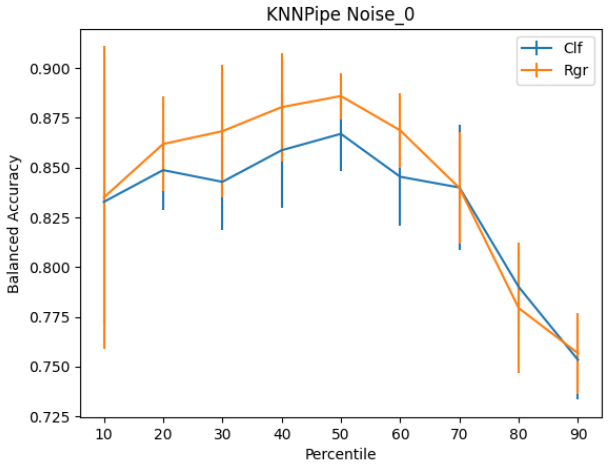


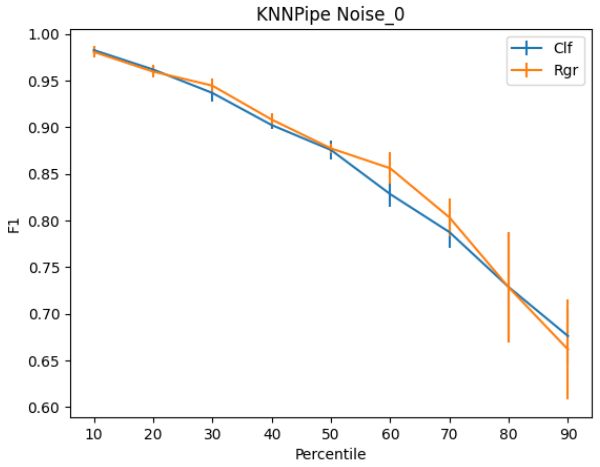


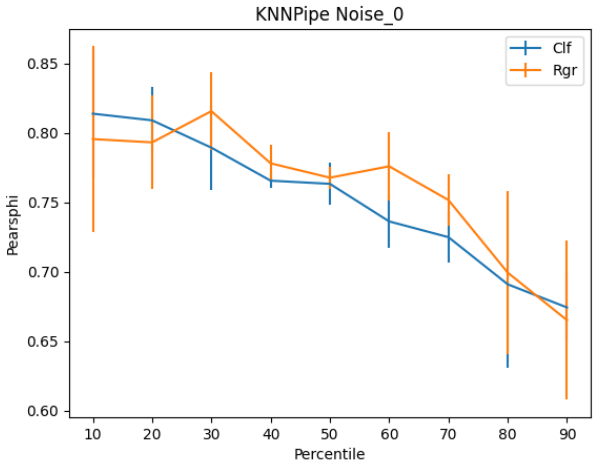


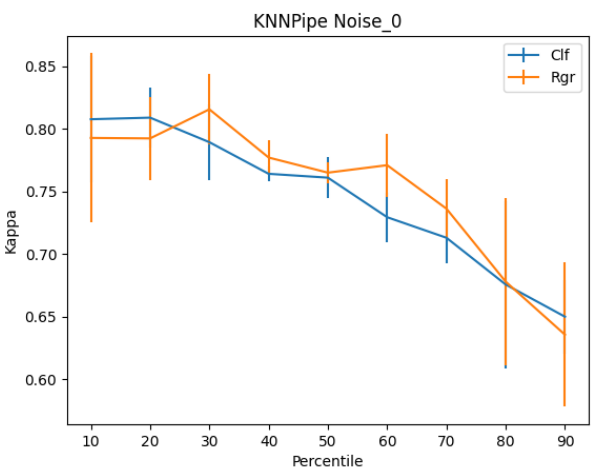


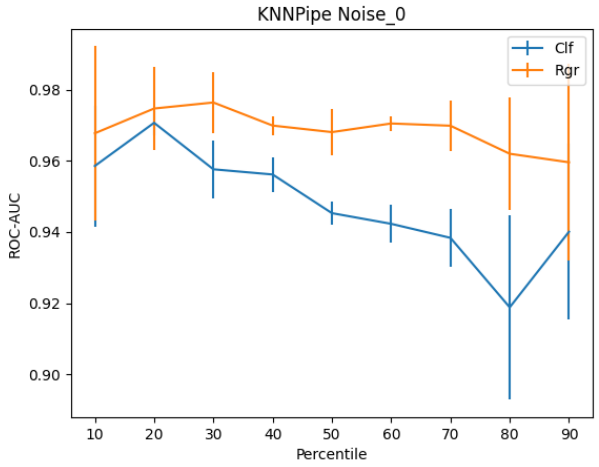
Multiple Metrics – *G298Atom + KNN*

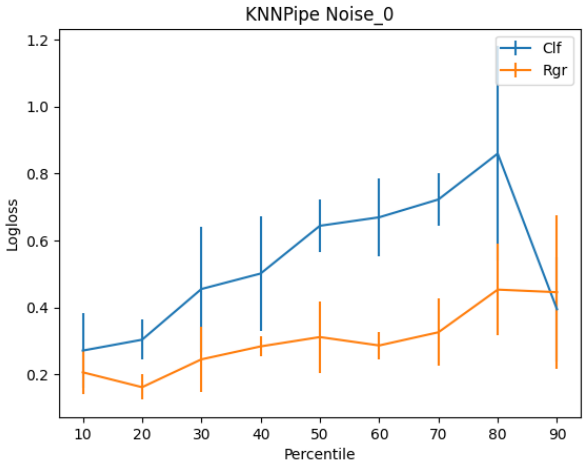


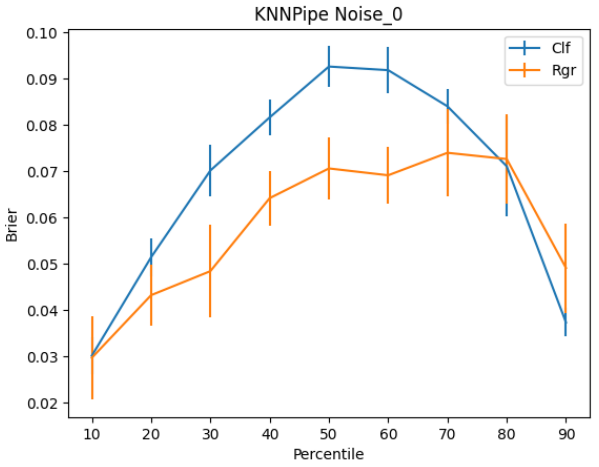




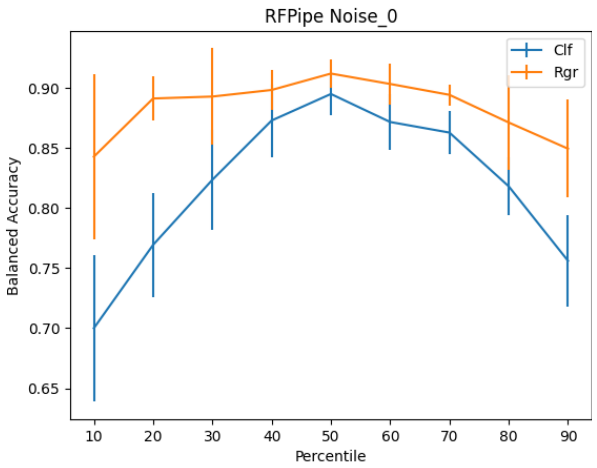


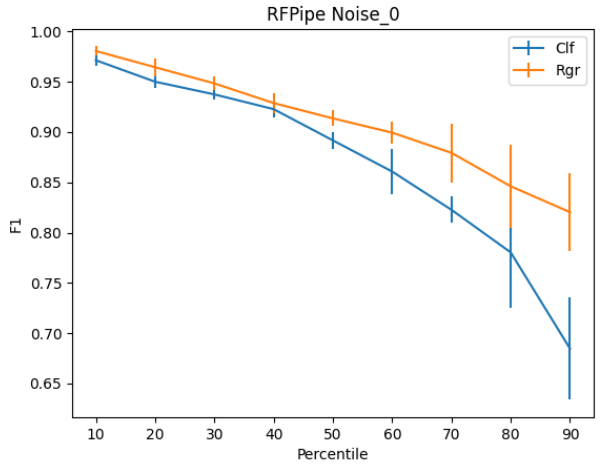


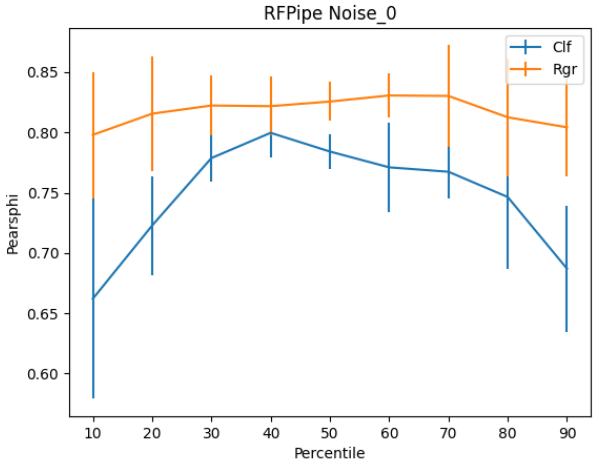


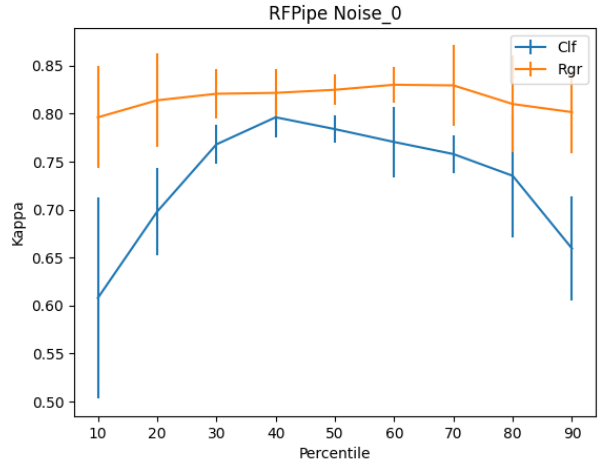


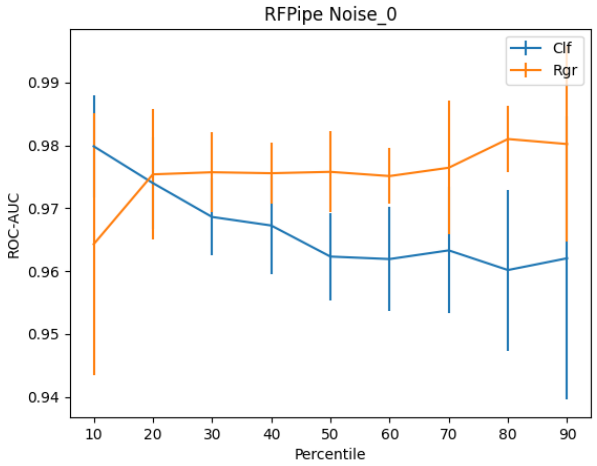
Multiple Metrics – *G298Atom + RF*

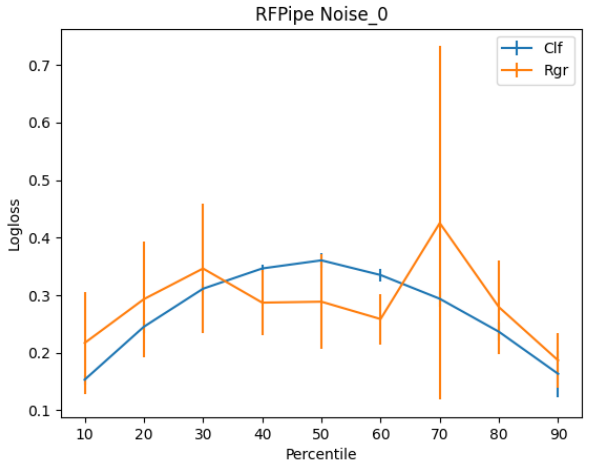


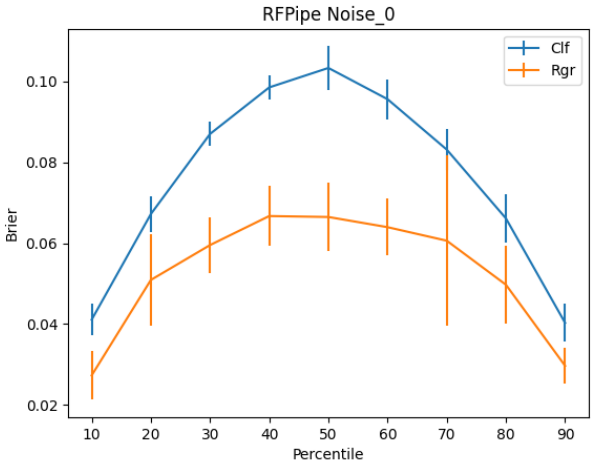




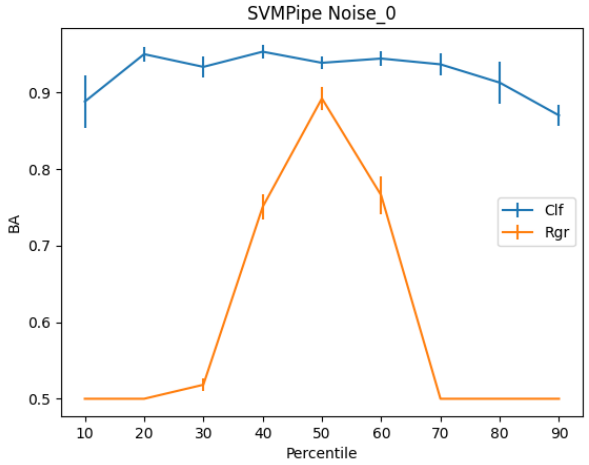


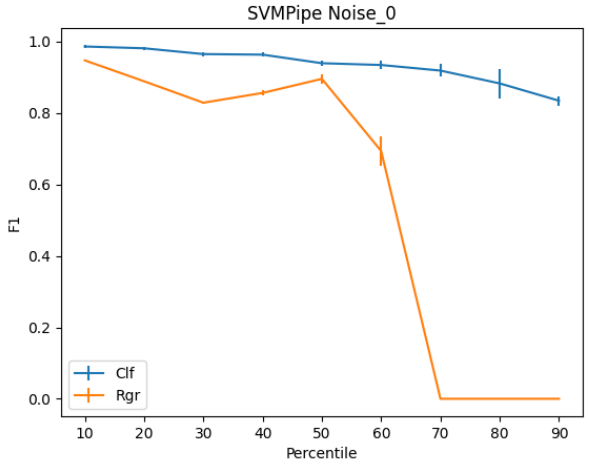


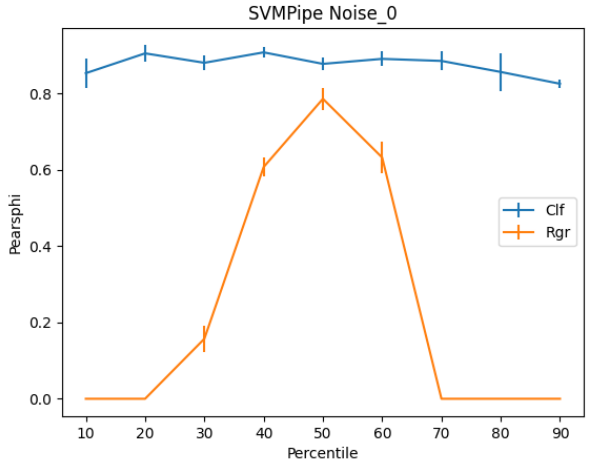


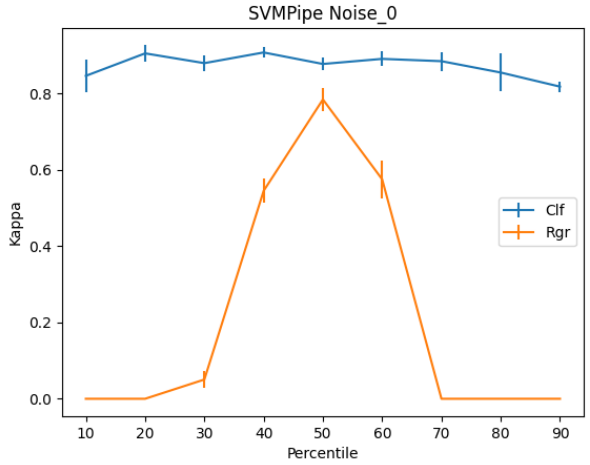


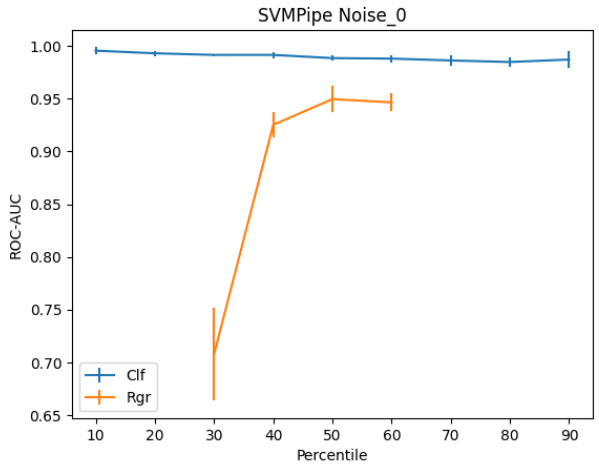
Multiple Metrics – *G298Atom + SVM*

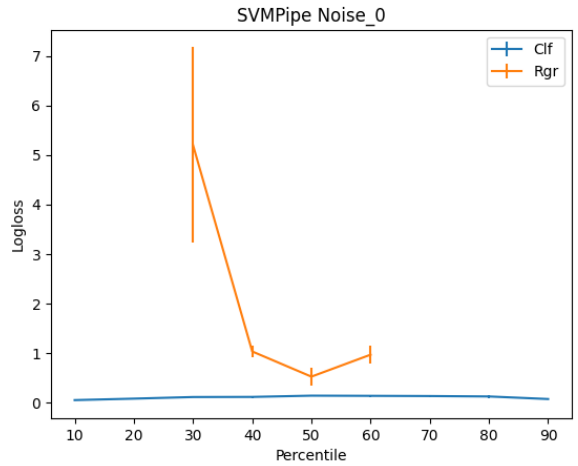


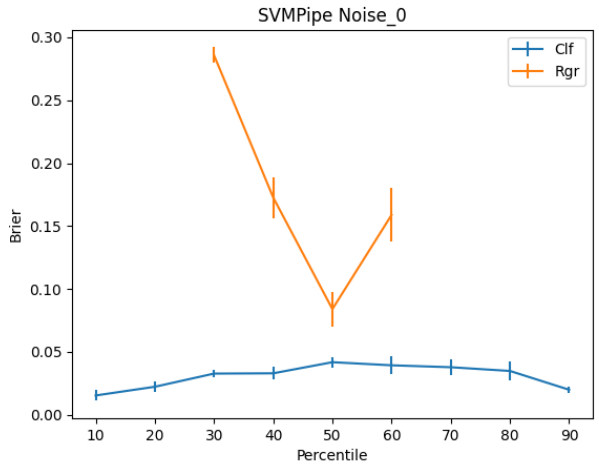












Aggregated Results

G298Atom

	BA	F1	PearsPhi	Kappa	ROC-AUC	LogLoss	Brier
DT	-	-	-	-	R	R	R
KNN	-	-	-	-	R	R	R
RF	R	R	R	R	R	-	R
SVM	C	C	C	C	C	C	C

Solv

	BA	F1	PearsPhi	Kappa	ROC-AUC	LogLoss	Brier
DT	-	-	-	-	R	R	R
KNN	-	-	-	-	C	C	C
RF	-	-	-	-	C	C	C
SVM	-	-	-	-	C	C	C

Results Overview

- *C* = classification
- *R* = regression
- Whichever method has a statistically significant advantage for most percentiles
- - = no statistical difference between methods

Aggregated Results

Tox102

	BA	F1	PearsPhi	Kappa	ROC-AUC	LogLoss	Brier
DT	-	-	-	-	-	R	R
KNN	-	-	-	-	C	-	C
RF	-	-	-	-	-	C	C
SVM	-	-	-	-	C	C	C

Tox134

	BA	F1	PearsPhi	Kappa	ROC-AUC	LogLoss	Brier
DT	-	-	-	-	R	R	R
KNN	-	-	-	-	-	R	-
RF	-	-	-	-	C	C	C
SVM	-	-	-	-	C	C	C

Results Overview

- *C* = classification
- *R* = regression
- Whichever method has a statistically significant advantage for most percentiles
- - = no statistical difference between methods

Conclusions

Approach

- Categorization of continuous data is bad statistical practice. But does it affect the predictivity of models?
- By making predictions *before* (*regression*) and *after* (*classification*) categorizing a continuous dataset, we can explore how categorization affects model performance

Results

- Relative performance of classification and regression is dependent on:
  - Cutpoint
  - Algorithm
  - Dataset

- Probabilistic metrics are needed to distinguish performance for some of the datasets

Ongoing Work

Open Research Questions

- How will the following affect relative performance?
  - Exploring the remaining benchmark datasets
  - Applying different levels of noise to the datasets
  - Applying a *correlation filter* to feature variables
  - Implementing *PCA* with *variable number of components*

- Using probabilistic metrics with binarized regression predictions is a somewhat arbitrary process
  - Currently the workflow applies a logistic regression step to derive probabilities from the binarized regression values
  - Are there alternative methods?

- Is the complexity of our machine learning pipeline obscuring the fundamental statistical differences between predicting before and after categorization?

Acknowledgements

Mentor

Chris Grulke

Internal Manuscript Review

Charles Lowe

Richard Judson

Computational Chemistry and Cheminformatics Branch (CCCB)

PIs

Daniel Chang
Chris Grulke
Paul Harten
Todd Martin
Grace Patlewicz
Ann Richard
Dan Vallero
Antony Williams

Postdocs and SSCs

Matthew Boyce
Zachary Chiodini
Willysha Jenkins
Charles Lowe
Christian Ramsland
Gabriel Sinclair
Tia Tate

Tox102 and Tox134 Datasets

Katie Paul-Friedman

Supplemental Slides

Essential Literature

Irwin and McClelland, *Journal of Marketing Research*, **2003**, *40*, 366.

- Categorization always reduces effect size (R and R^2) between variables
- This effect holds for non-normal distributions
- This reduces power in *simple regression modeling*

MacCallum et al. *Psychological Methods*, **2002**, *7*, 19.

- Categorization always reduces effect size (R and R^2) between variables
- Derives the fundamental statistics
- Debunks common justifications for categorization
- Concludes that categorization is rarely defensible

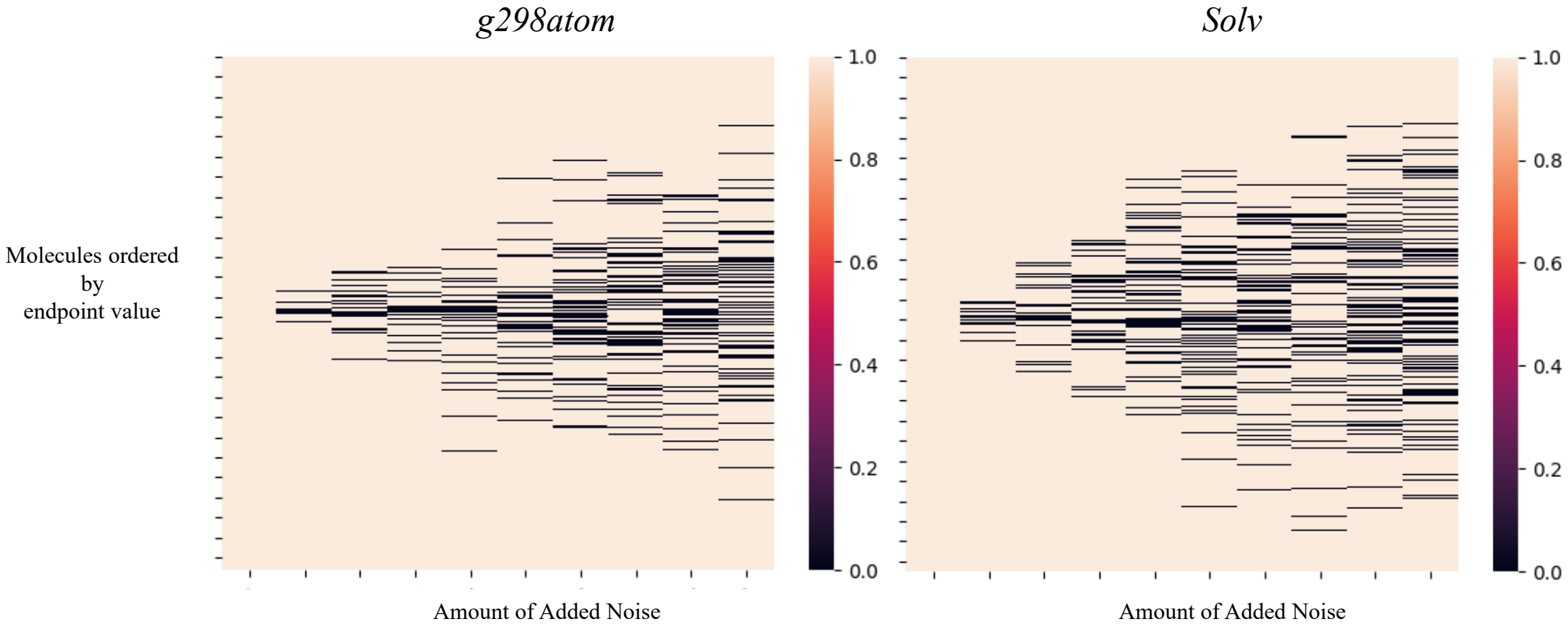
Royston et al. *Statist. Med.*, **2006**, *25*, 127.

Table II. Quantifying the loss of information in two cutpoint models for the PBC data, compared with model 3 in which continuous variables were retained as continuous.

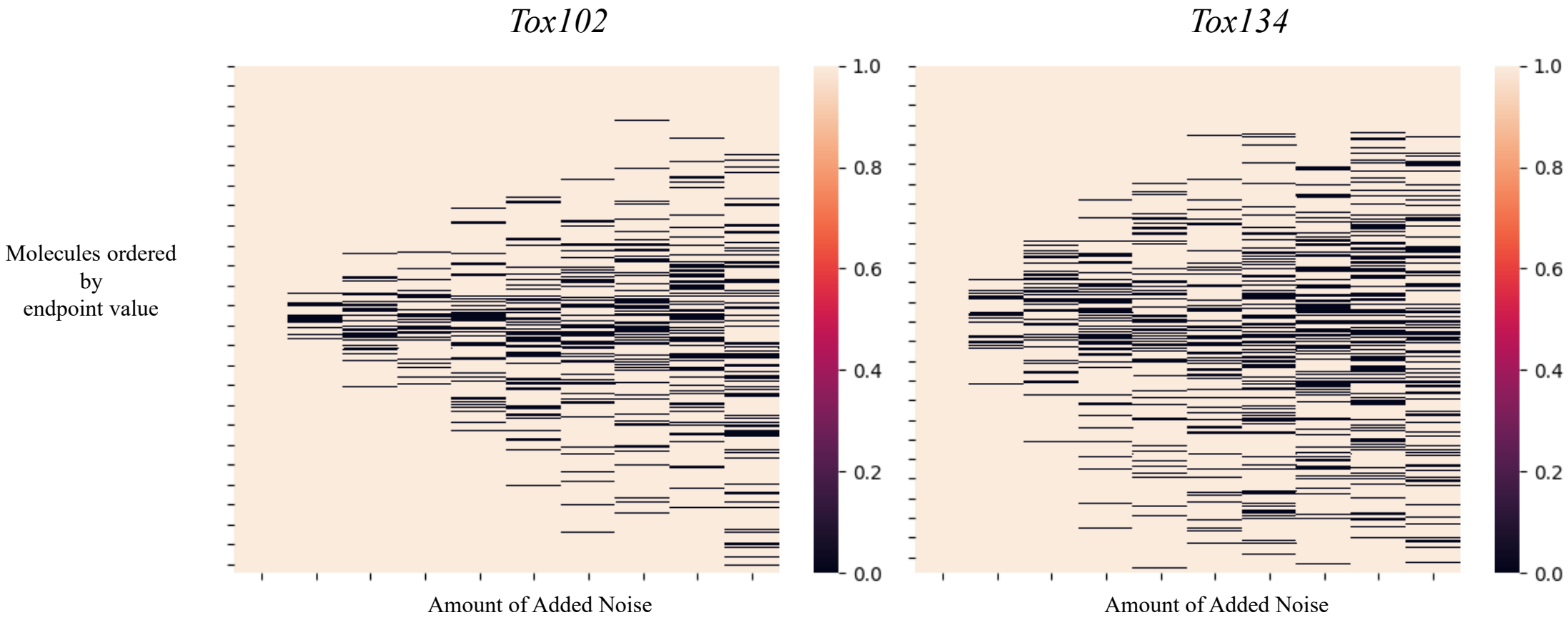
Measure	Model 1 Median cutpoint	Model 2 'Optimal' cutpoint	Model 3 Continuous
Model χ^2	94.6	99.2	136.8
c-index	0.778	0.774	0.814
D	1.91	2.02	2.55
R^2_D	0.465	0.494	0.608

- Categorizing continuous data results in less predictive models

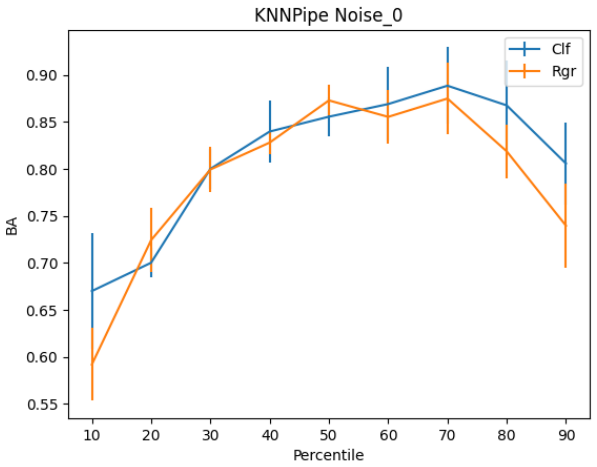
The Effect of Noise on Class Split

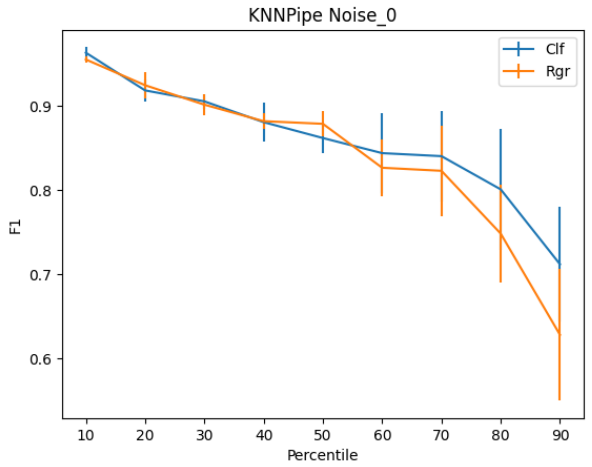


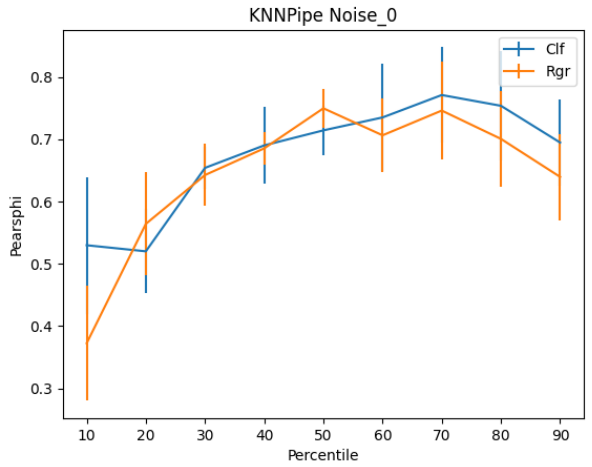
The Effect of Noise on Class Split

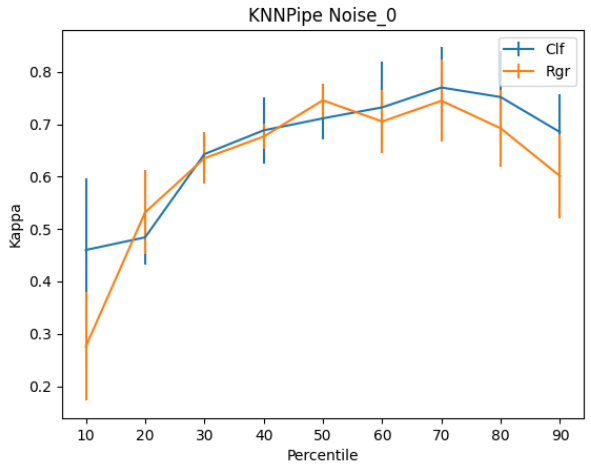


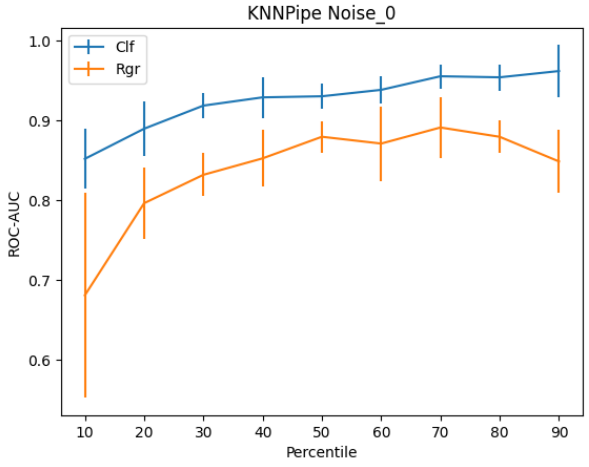
Multiple Metrics – *Solv + KNN*

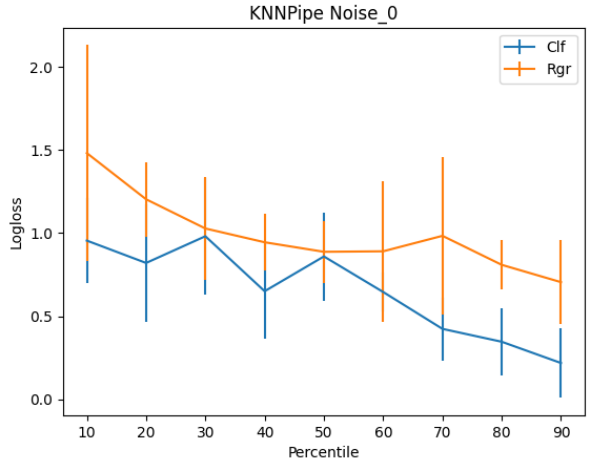


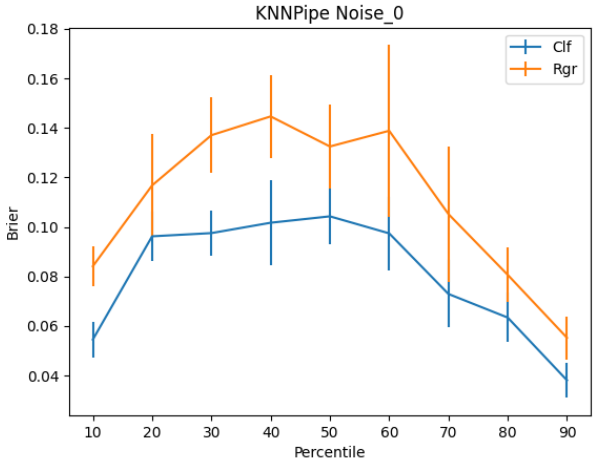




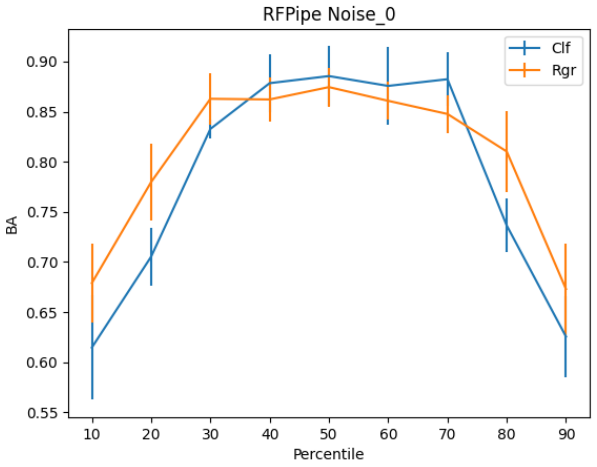


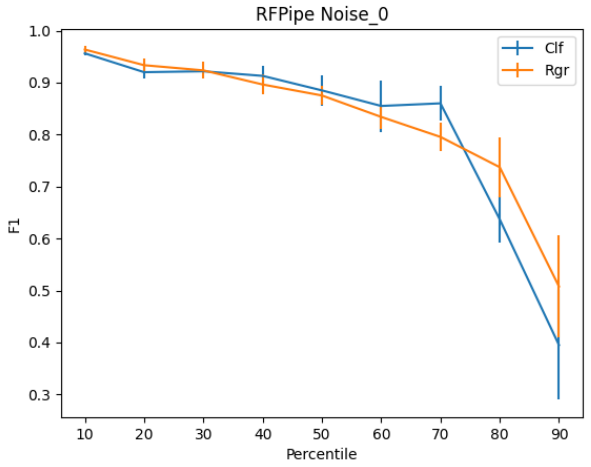


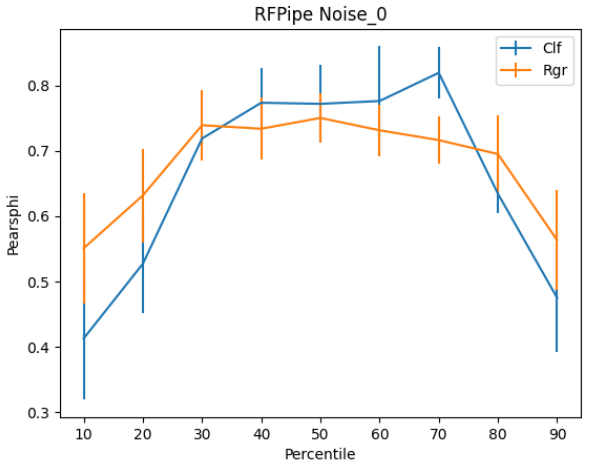


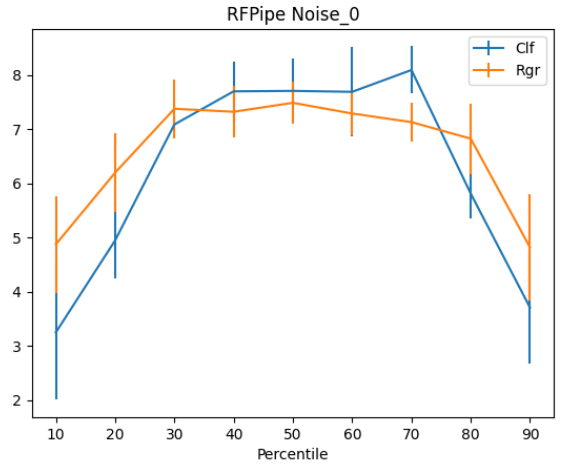


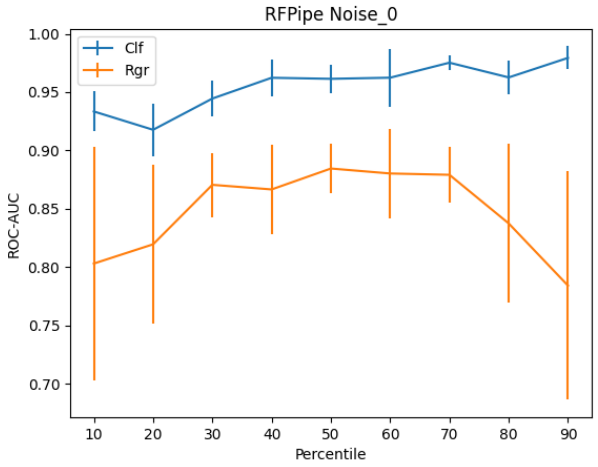
Multiple Metrics – *Solv + RF*

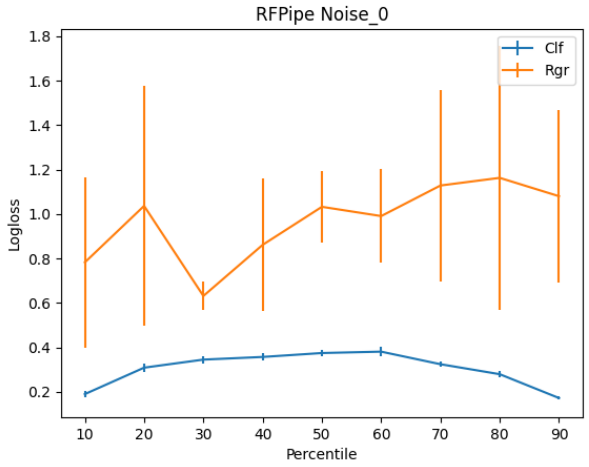


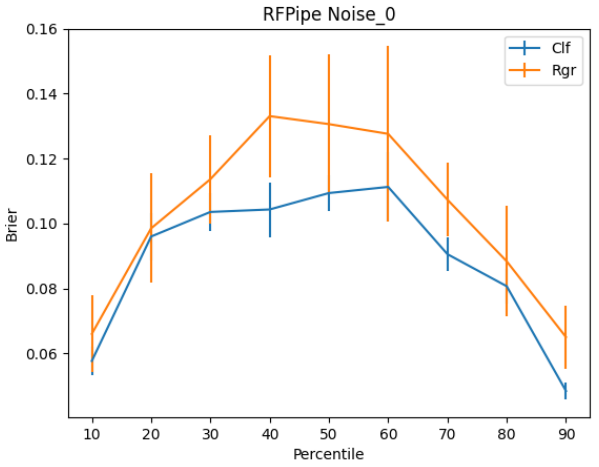




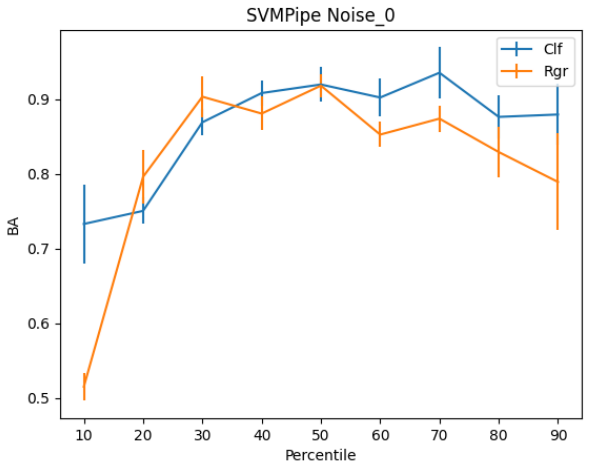


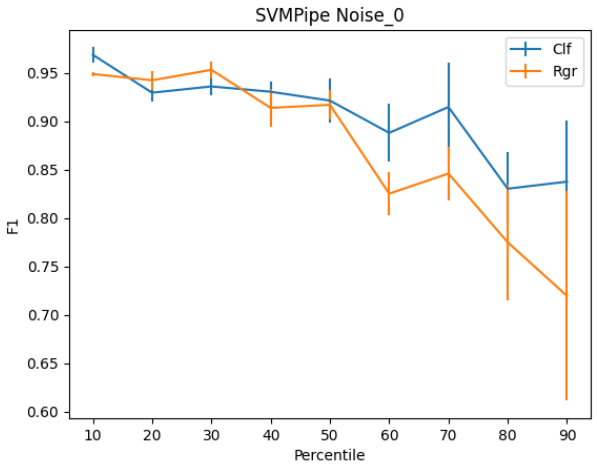


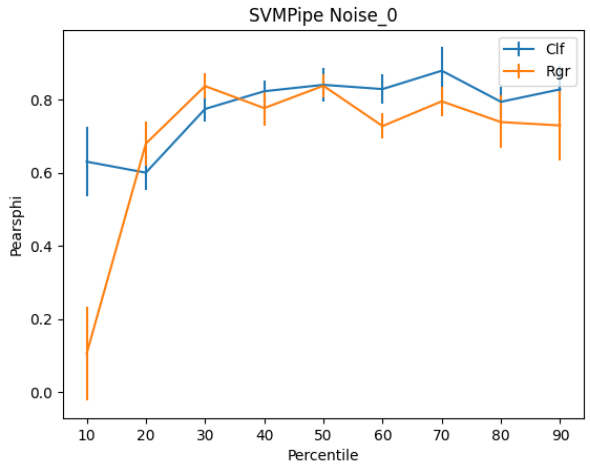


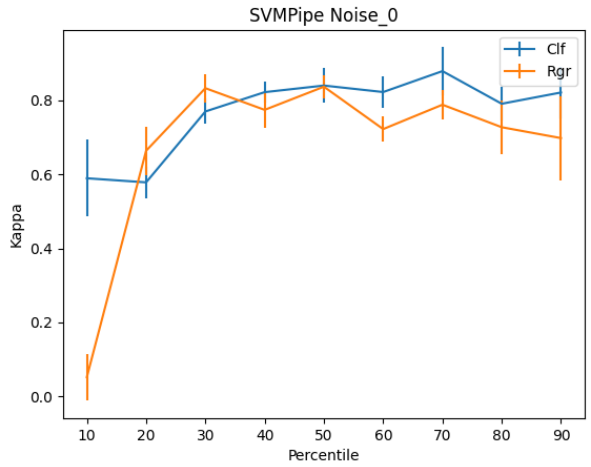


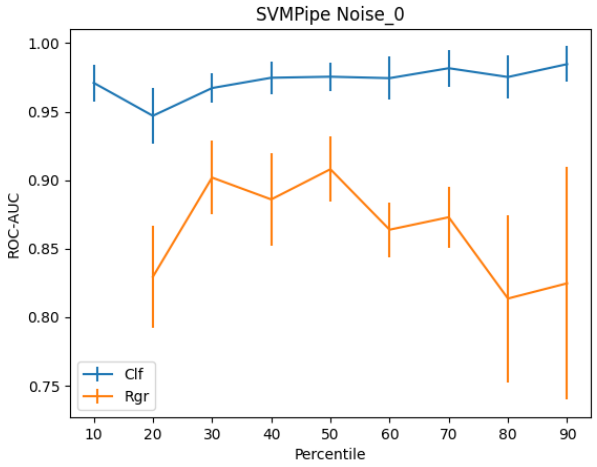
Multiple Metrics – *Solv + SVM*

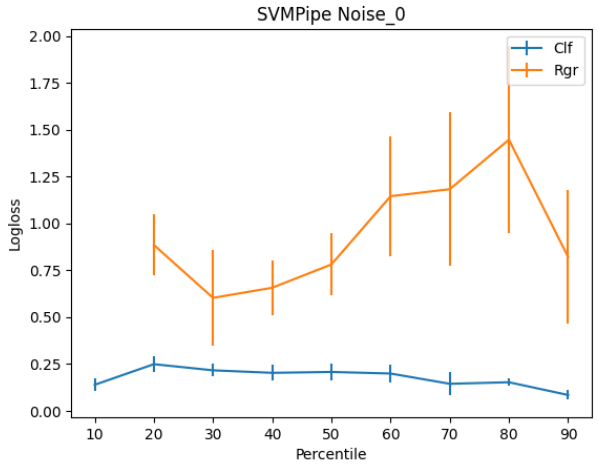


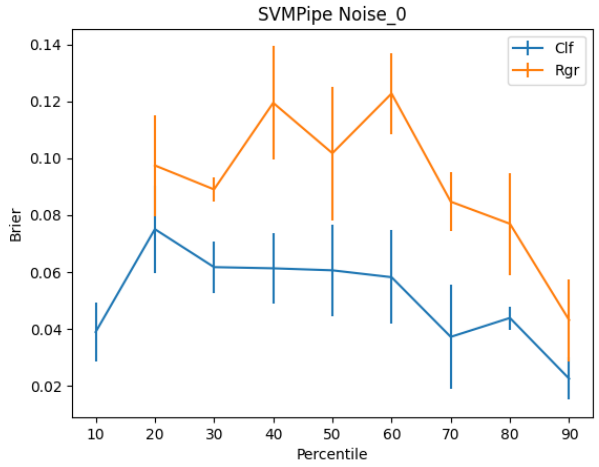




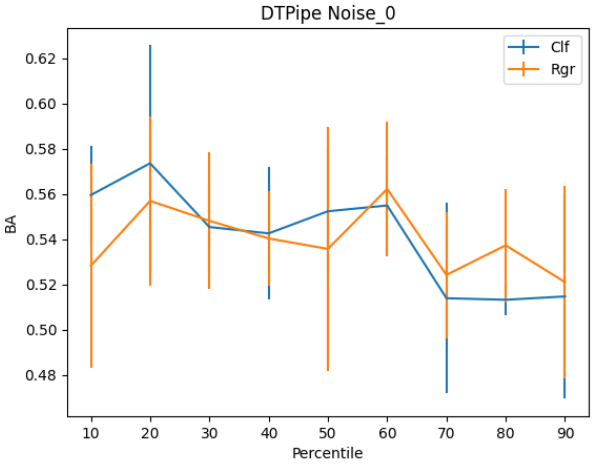


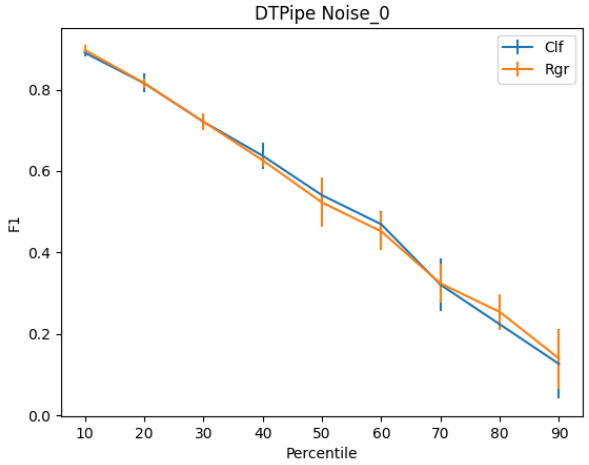


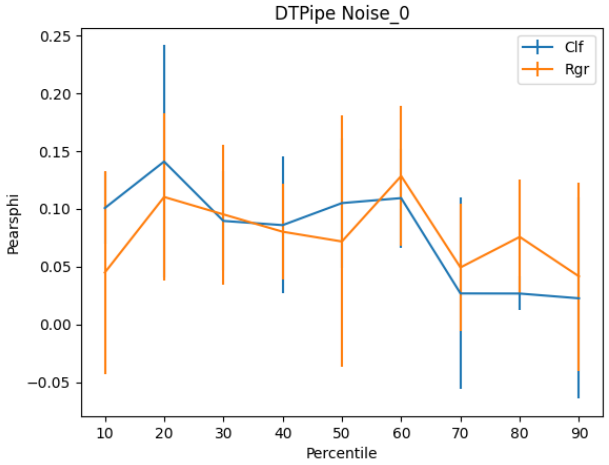


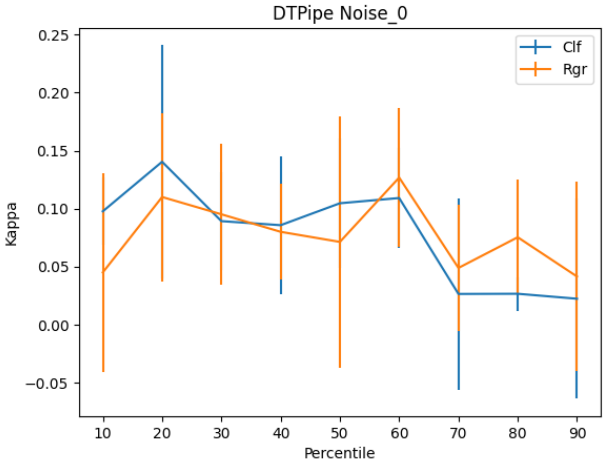


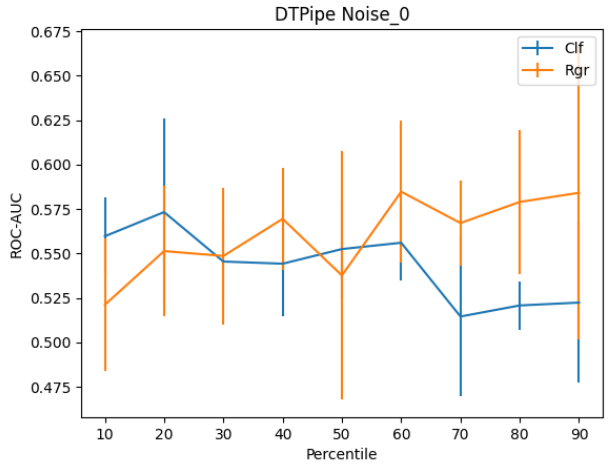
Multiple Metrics – *Tox102 + DT*

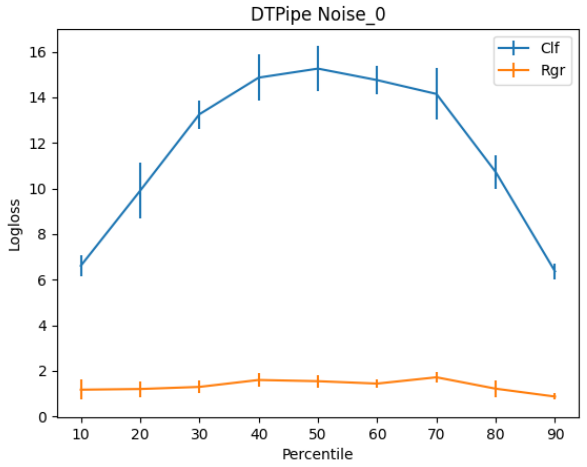


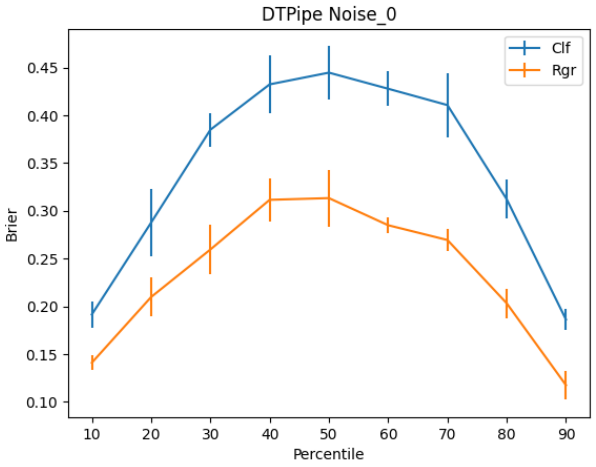




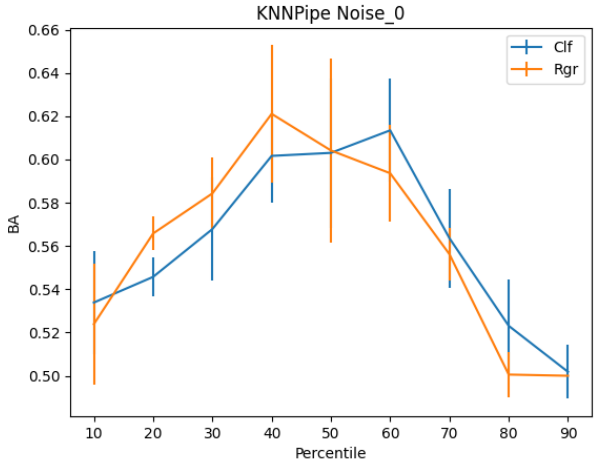


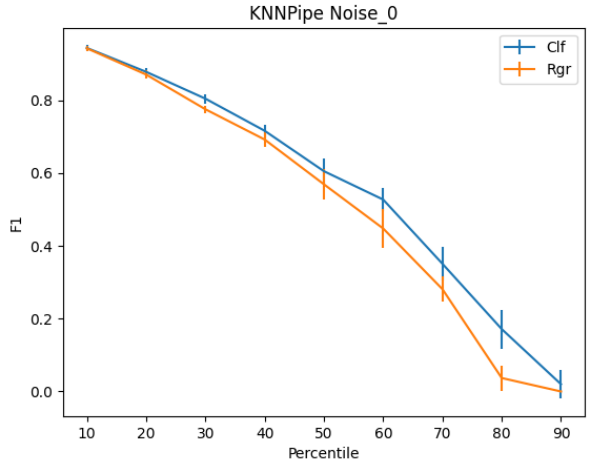


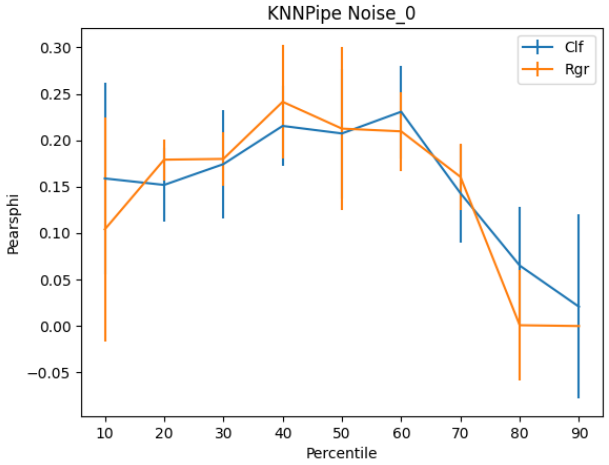


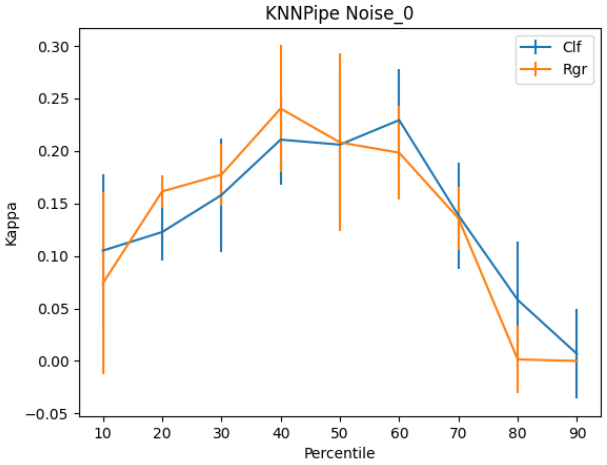


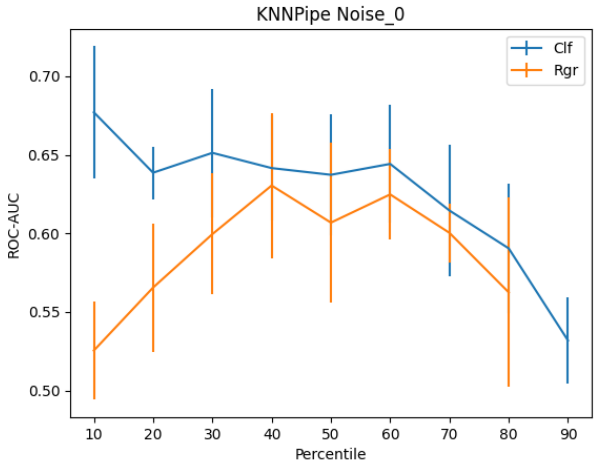
Multiple Metrics – *Tox102 + KNN*

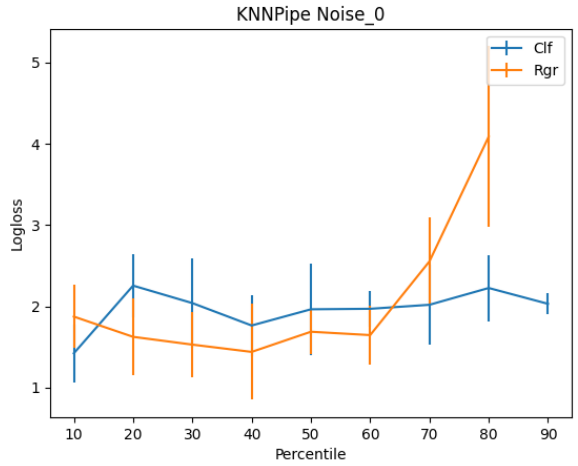


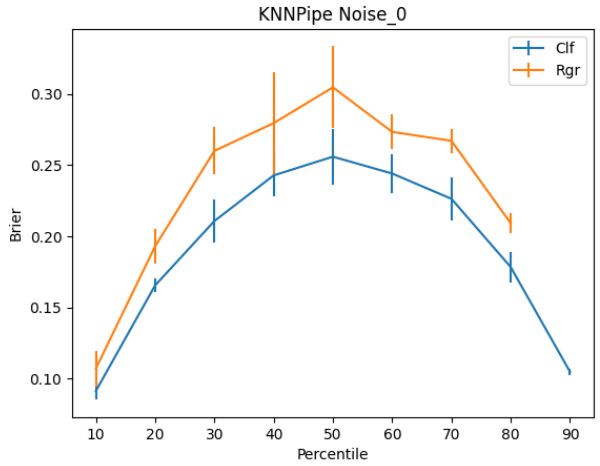




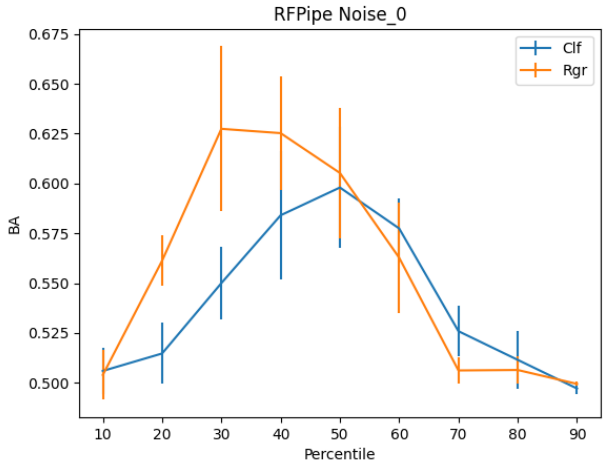


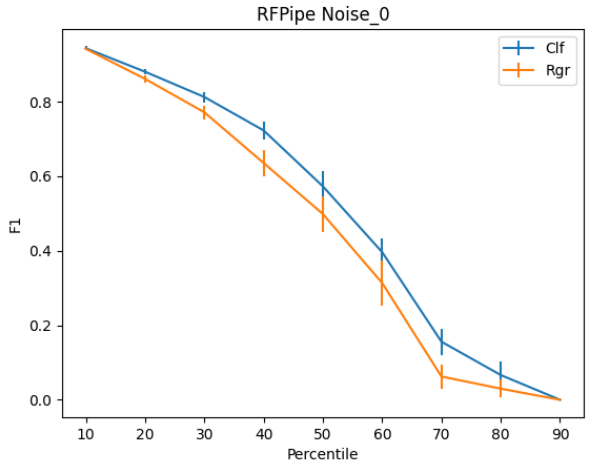


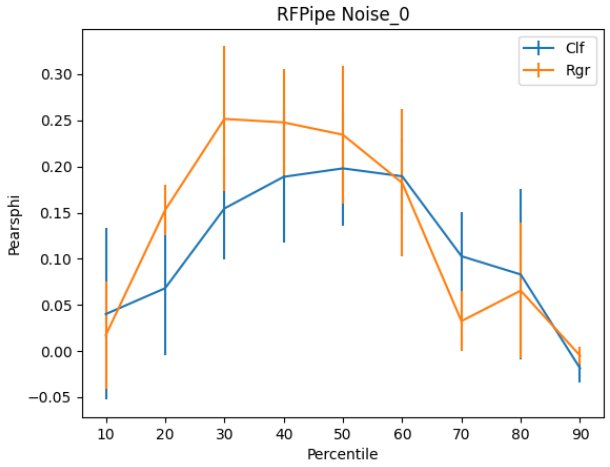


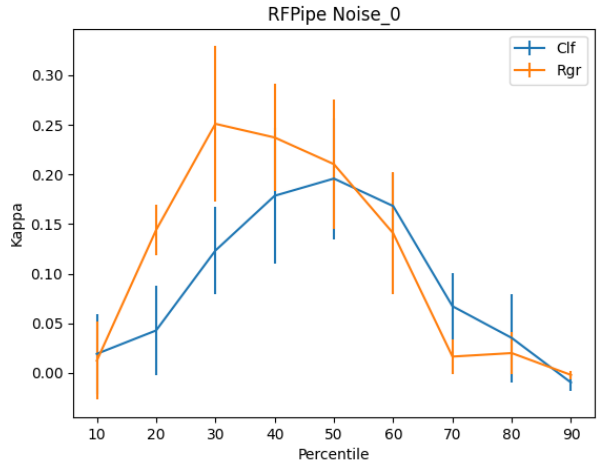


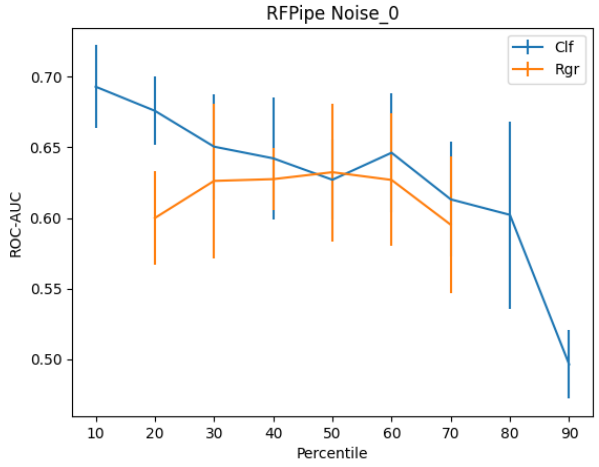
Multiple Metrics – *Tox102 + RF*

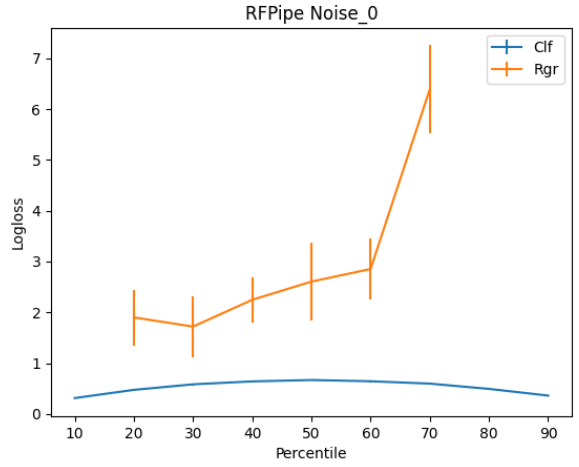


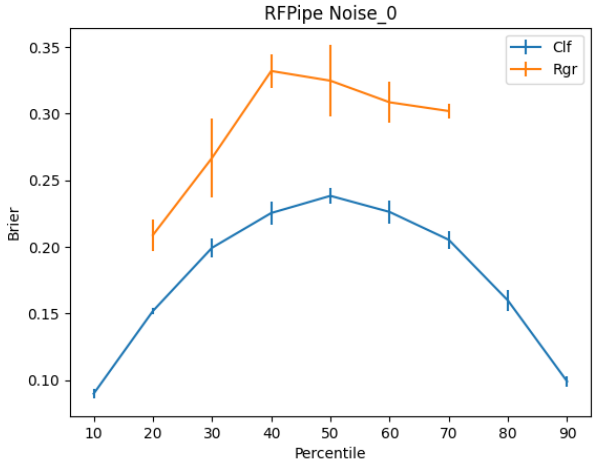




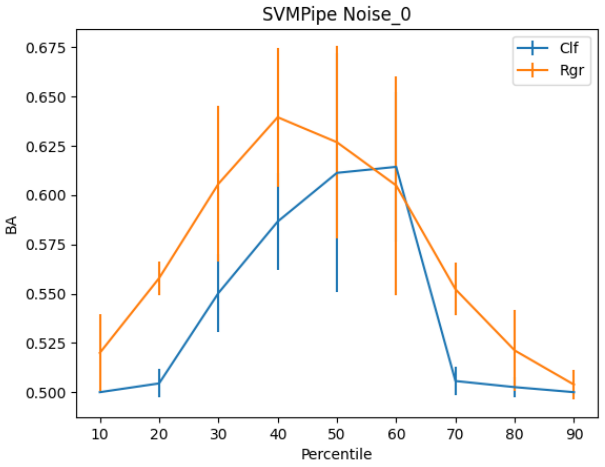


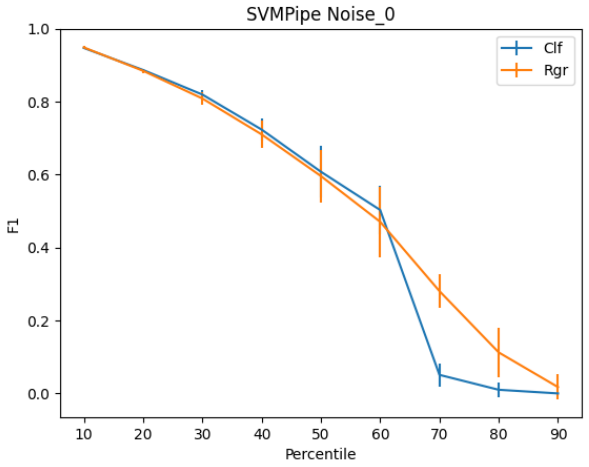


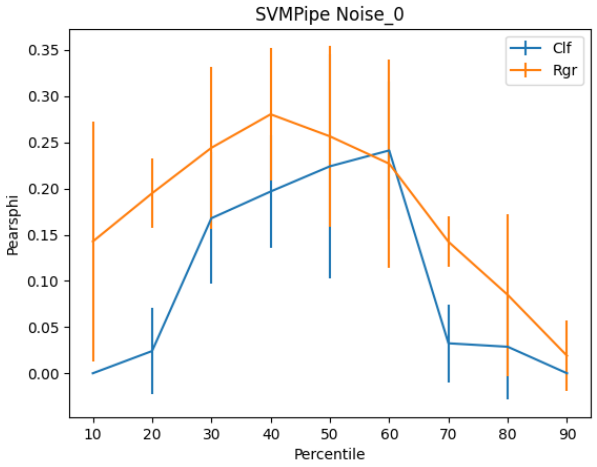


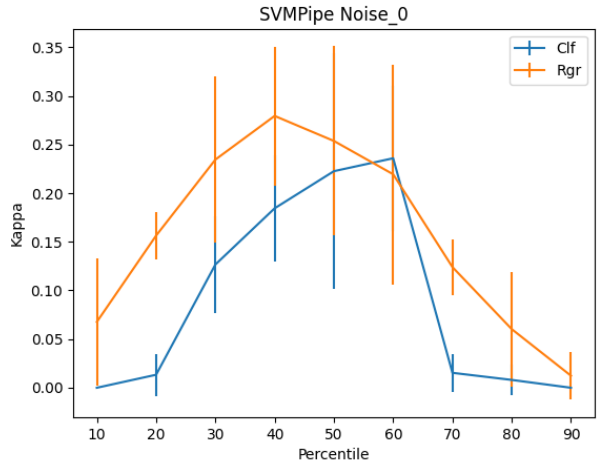


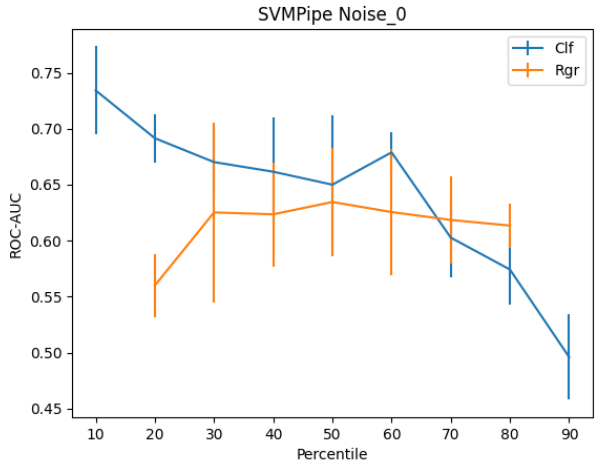
Multiple Metrics – *Tox102 + SVM*

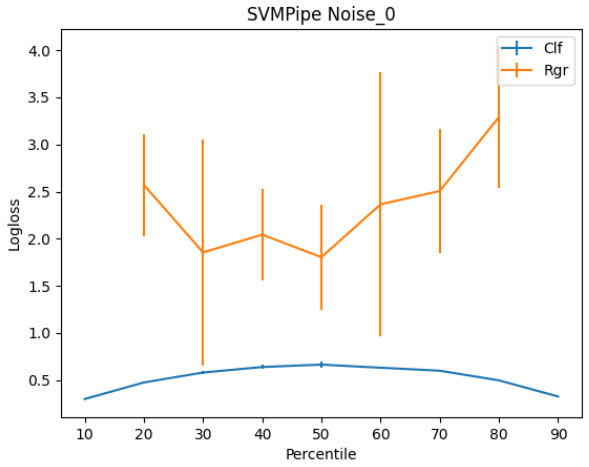


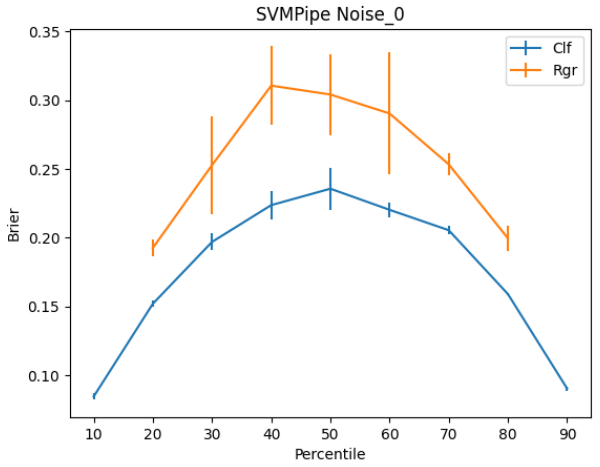




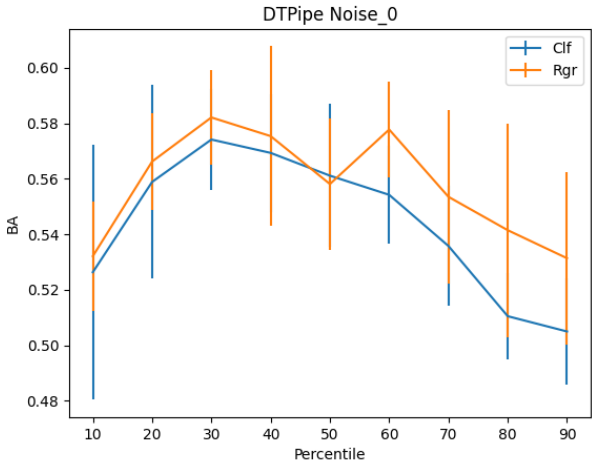


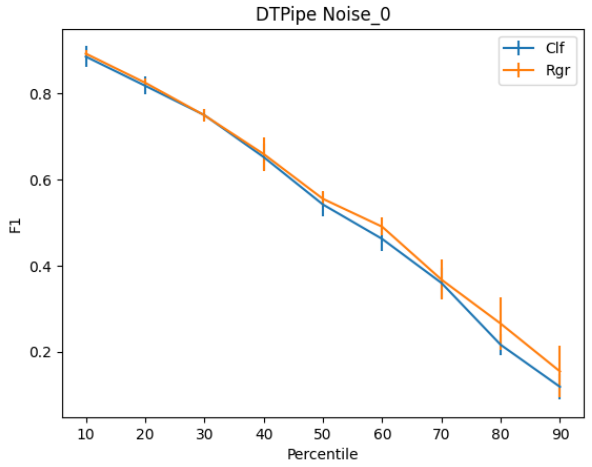


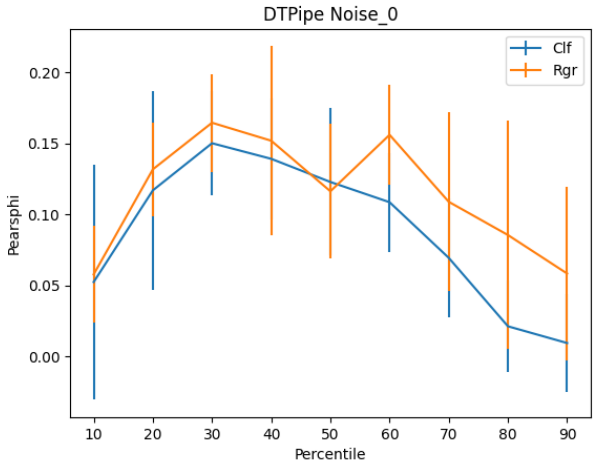


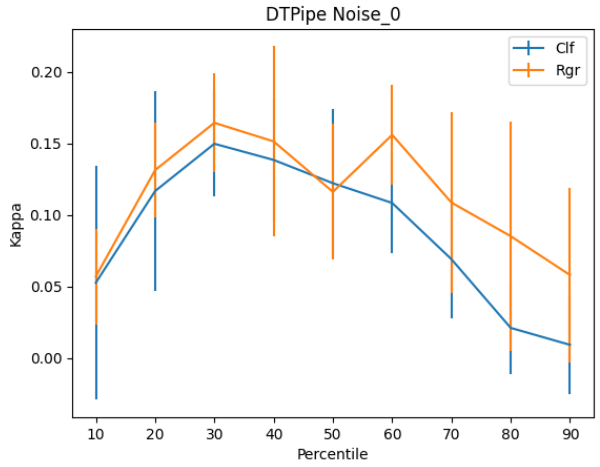


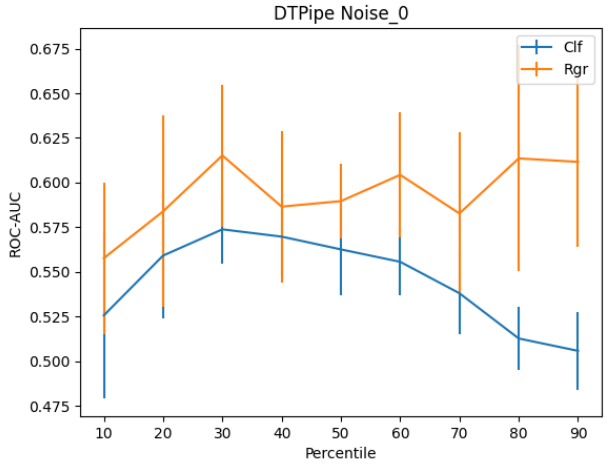
Multiple Metrics – *Tox134 + DT*

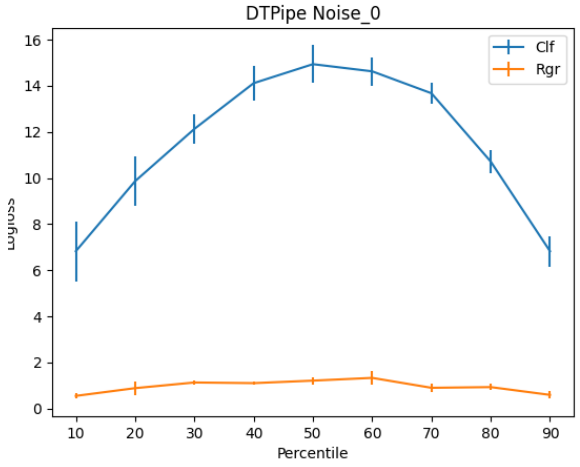


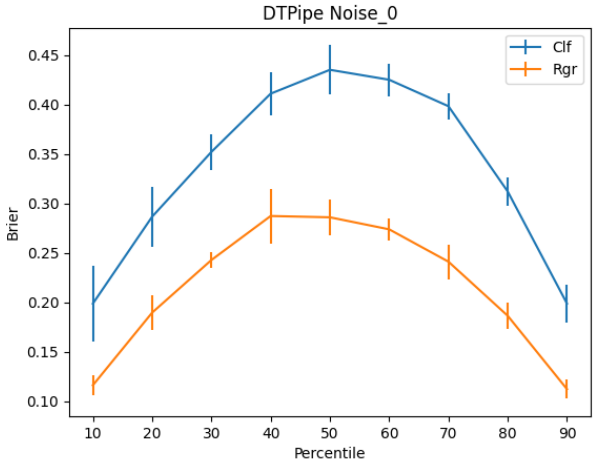




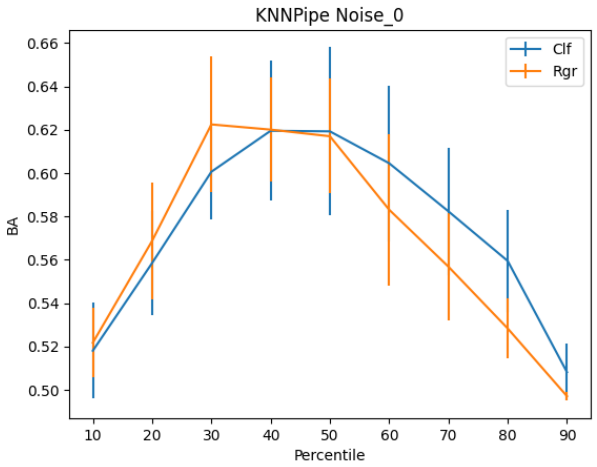


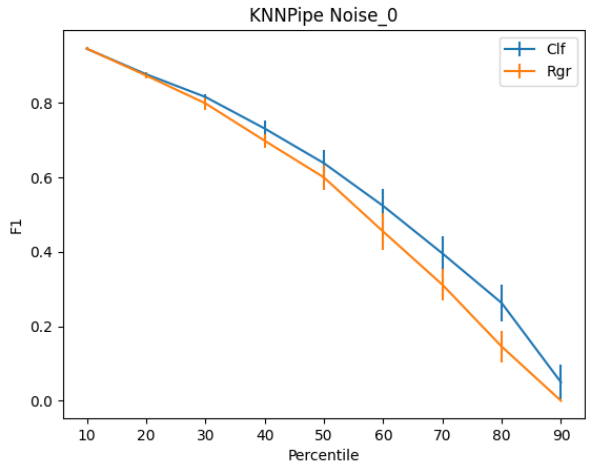


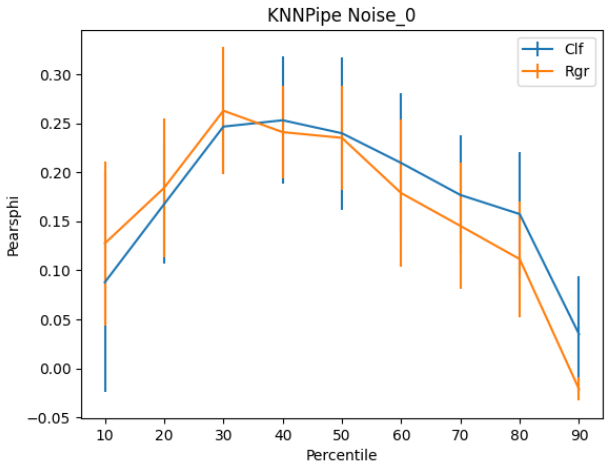


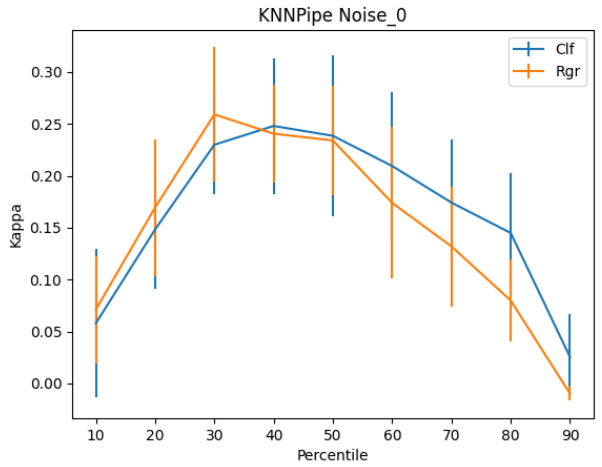


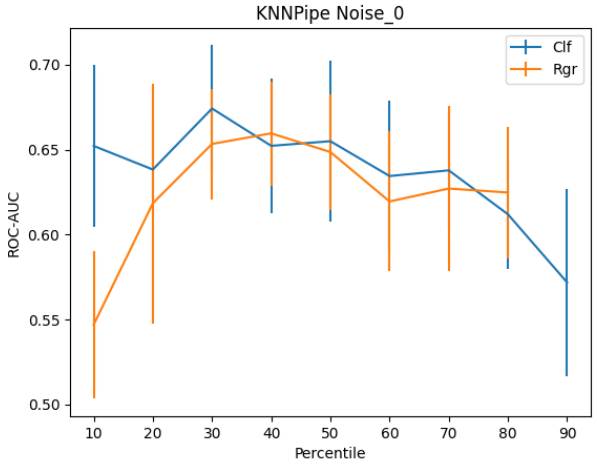
Multiple Metrics – *Tox134 + KNN*

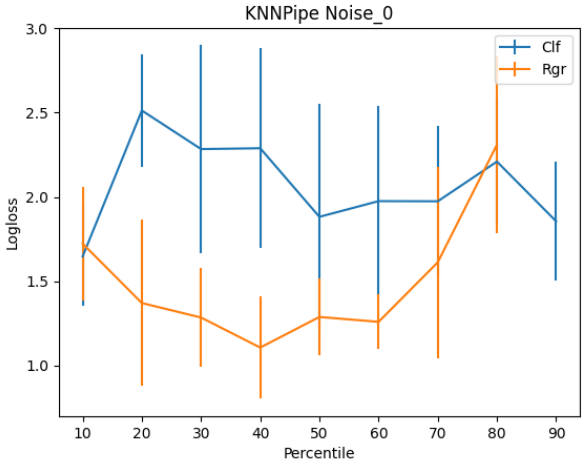


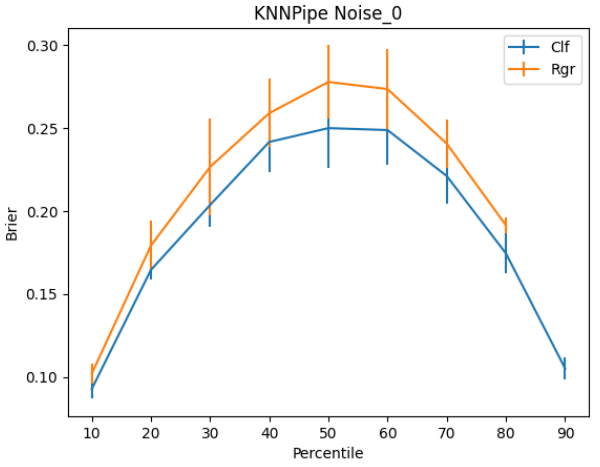




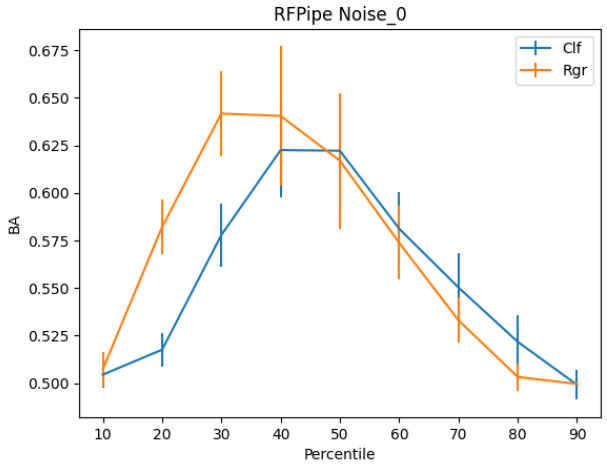


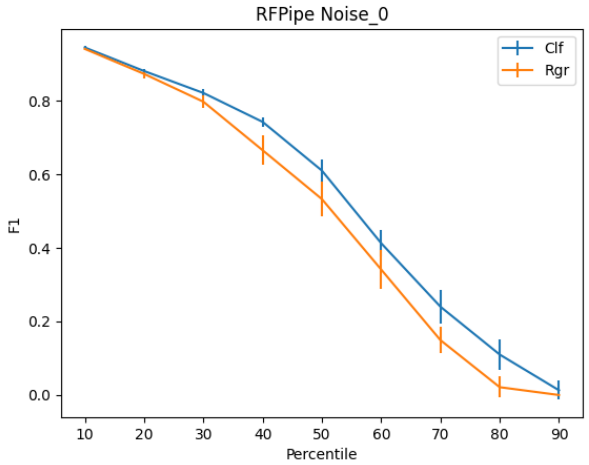


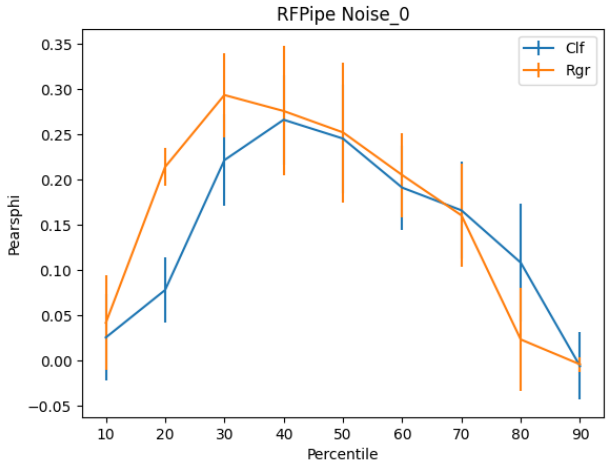


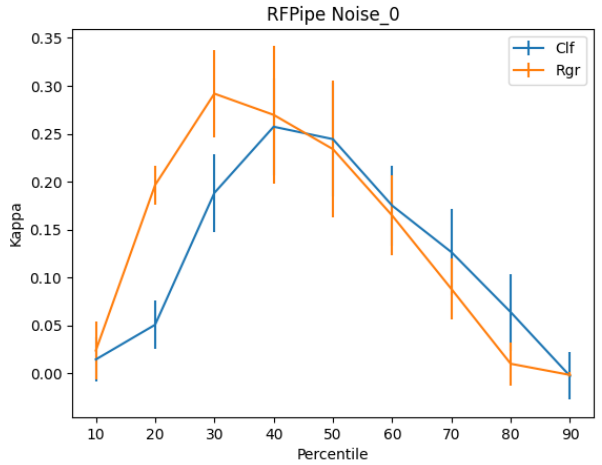


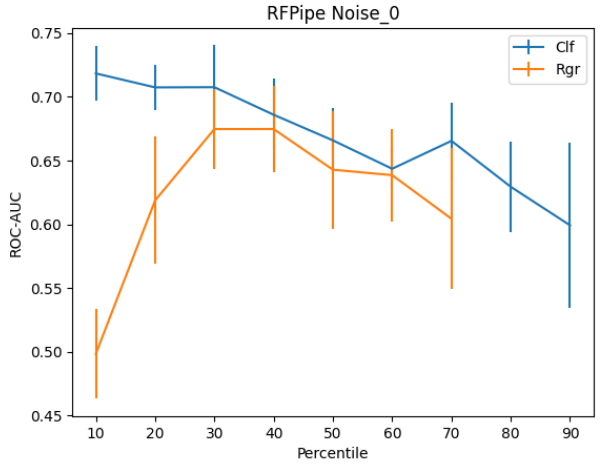
Multiple Metrics – *Tox134 + RF*

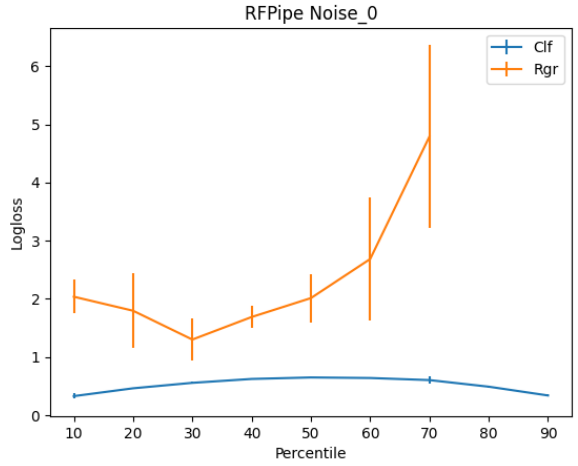


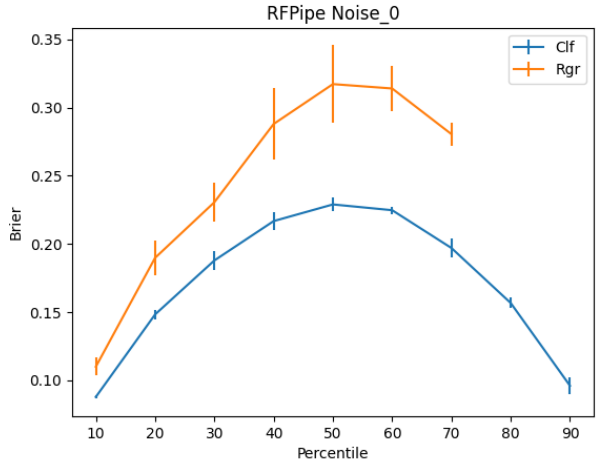




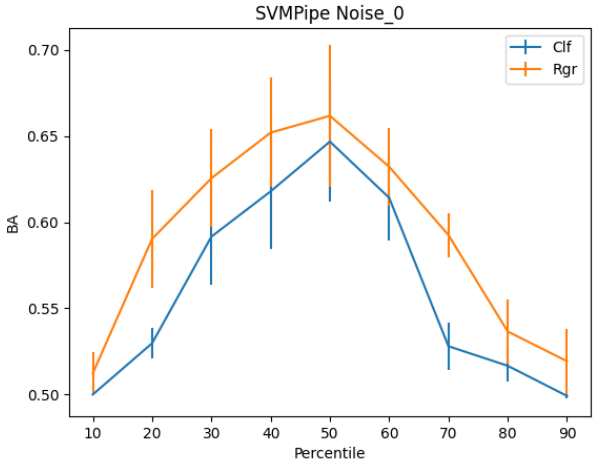


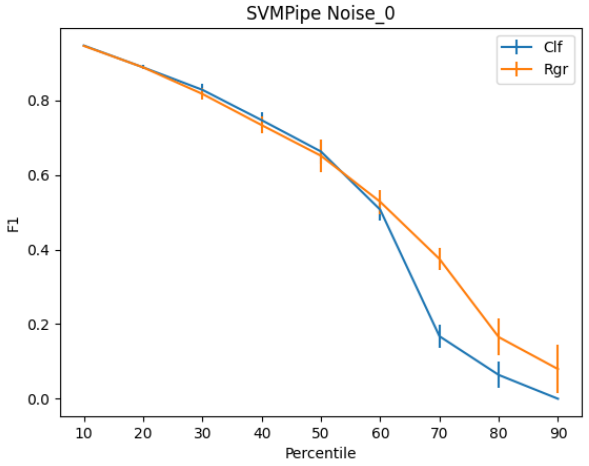


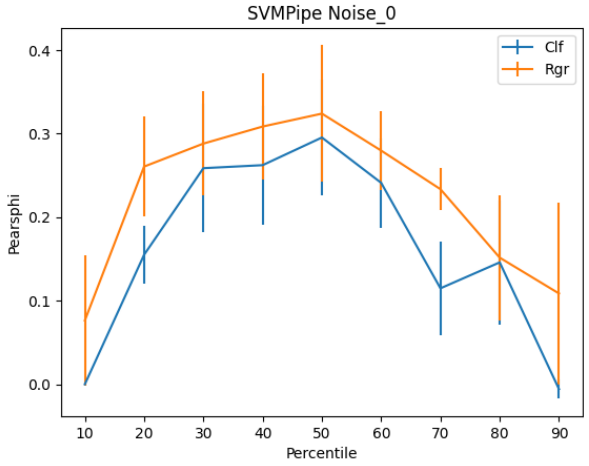


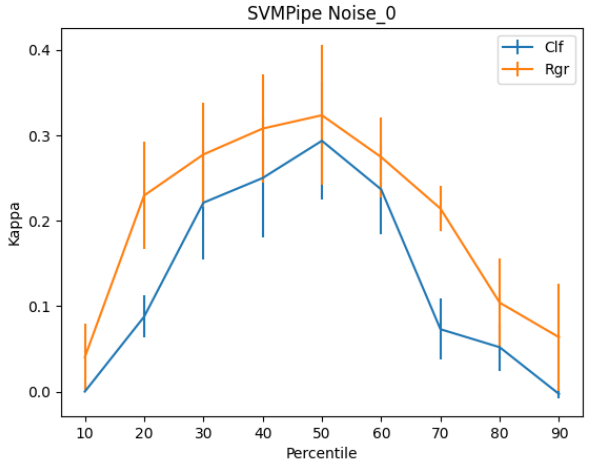


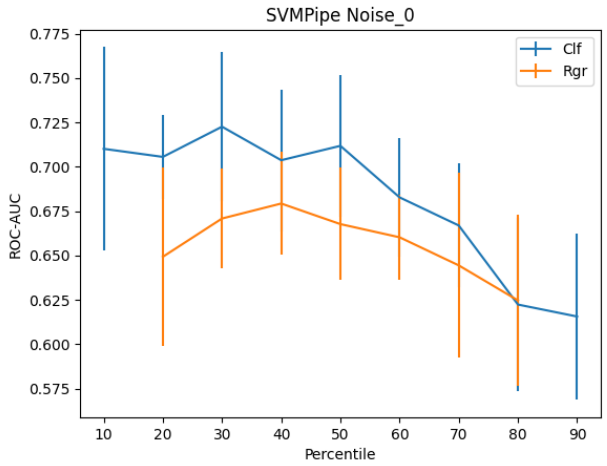
Multiple Metrics – *Tox134 + SVM*

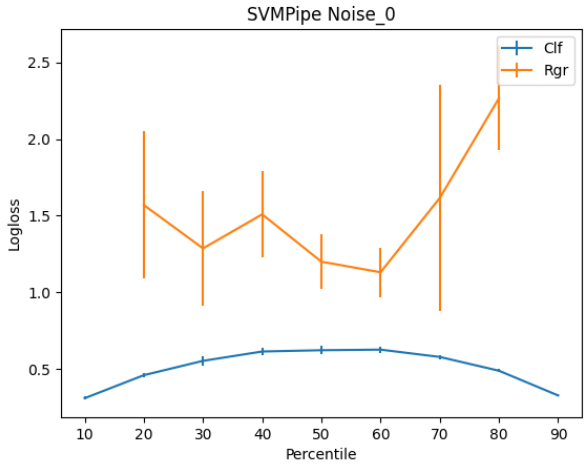


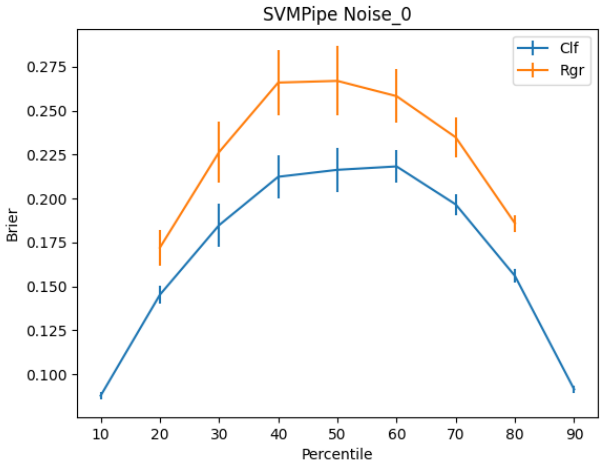












Feature Variables

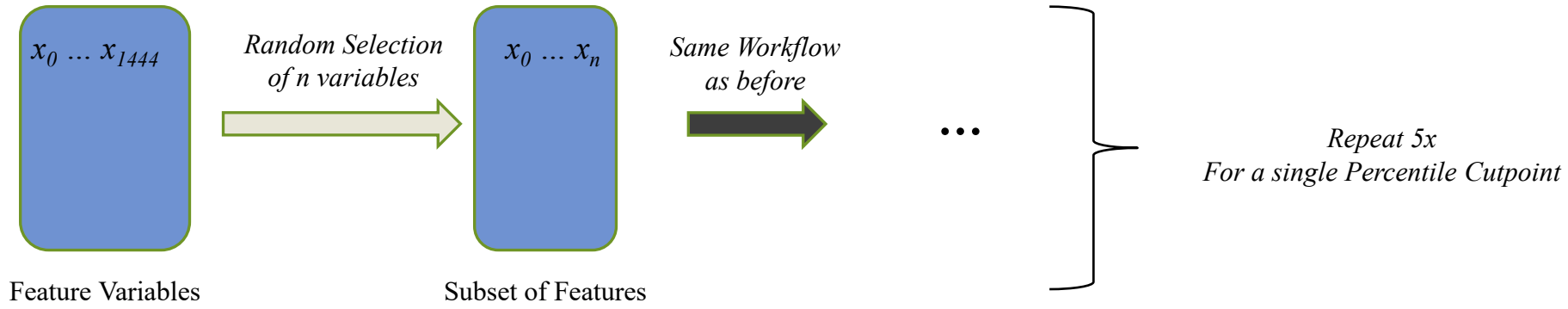
Hypothesis

The statistics show that when there are few *feature variables* (~ < *3)*, categorization of a continuous variable leads to clear loss of statistical power.

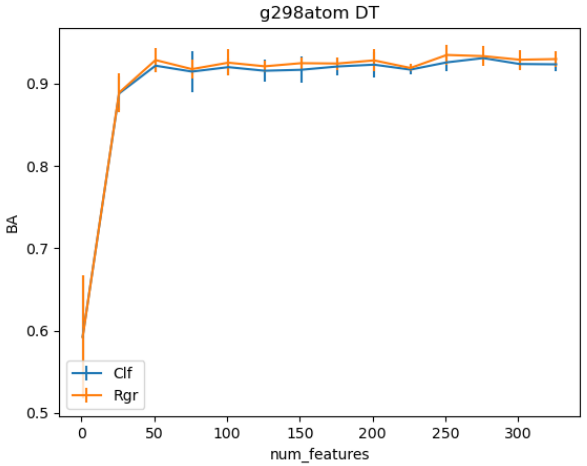
Our models contain 1400 *feature variables* in the standard workflow, and 100 *feature variables* after Principal Component Analysis (*PCA*).

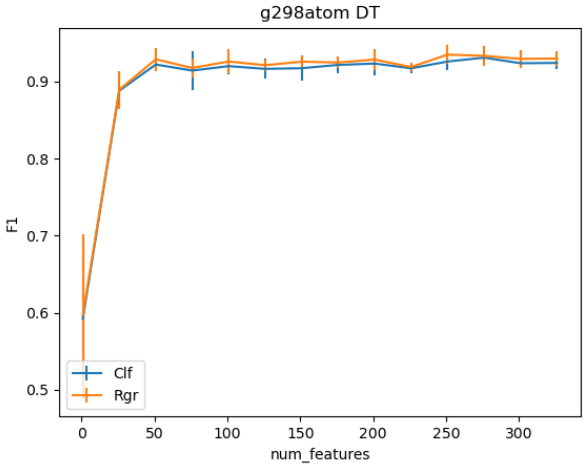
What is the relationship between the number of *feature variables* and the relative performance of the classification and regression methods?

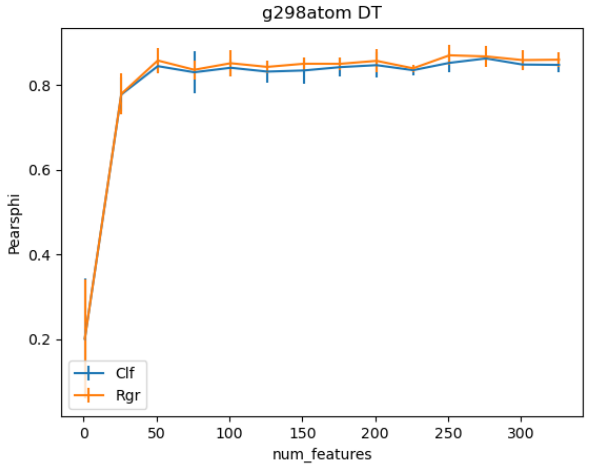
Approach

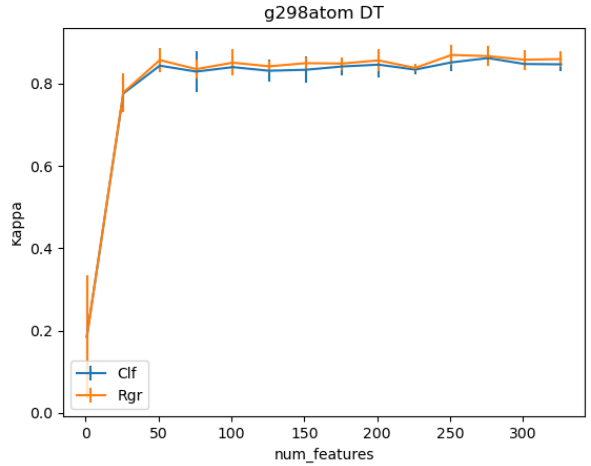


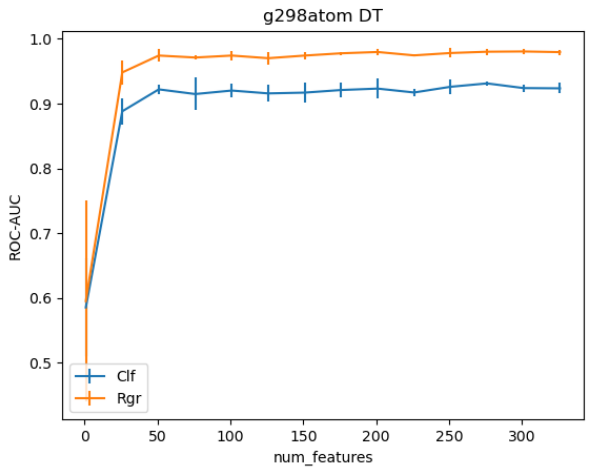
Feature Variable Results - *G298Atom + 50%*

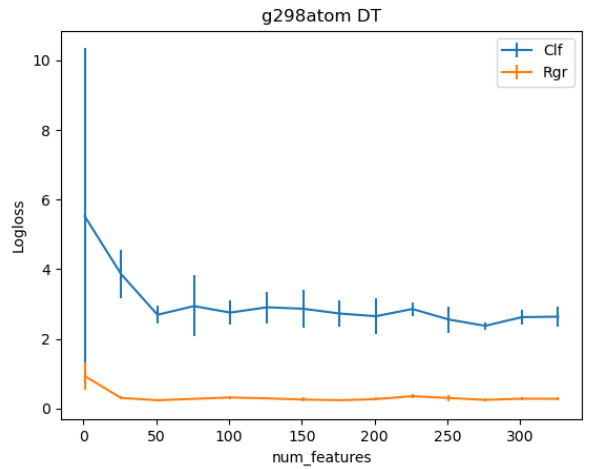


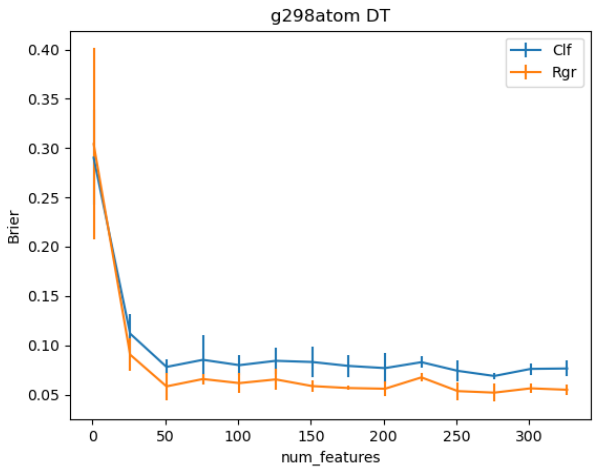












Feature Variable Results – *Solv + 50%*