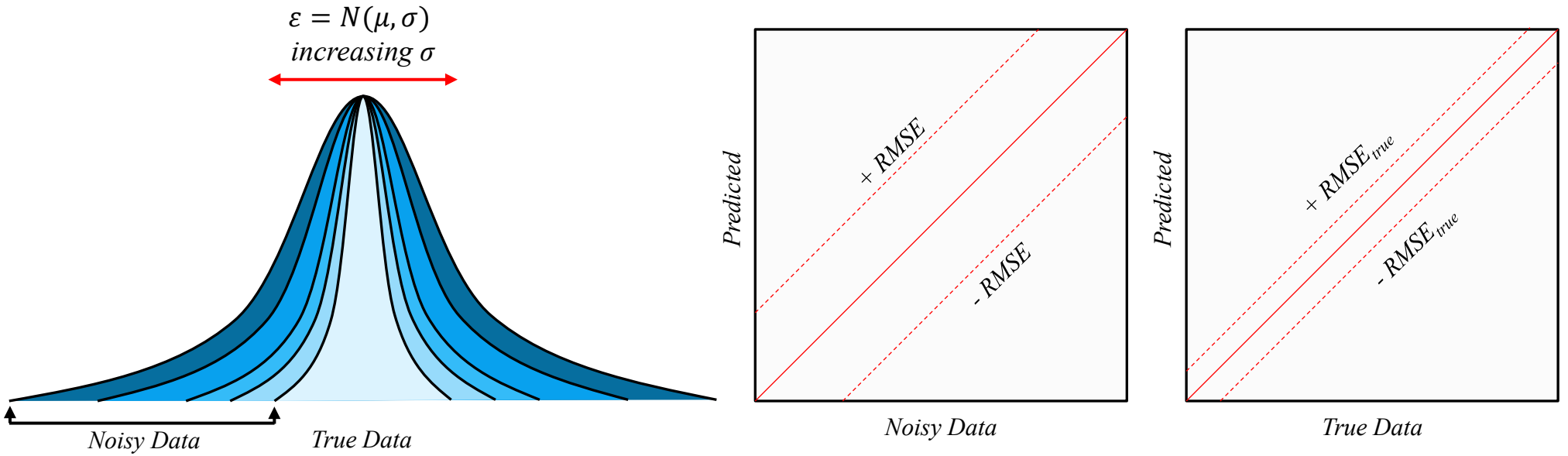
Determining the Predictive Limit of QSAR Models



Scott Kolmar

ORCID: 0000-0002-7797-700X

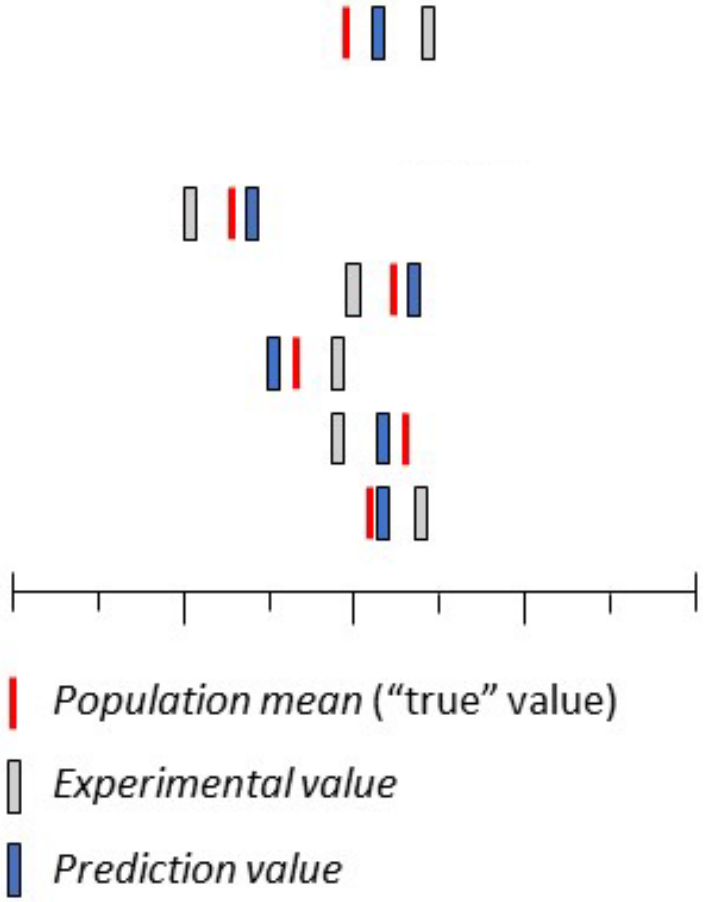
June 9th, 2021

US EPA

ORD-CCTE-CCED-CCCB

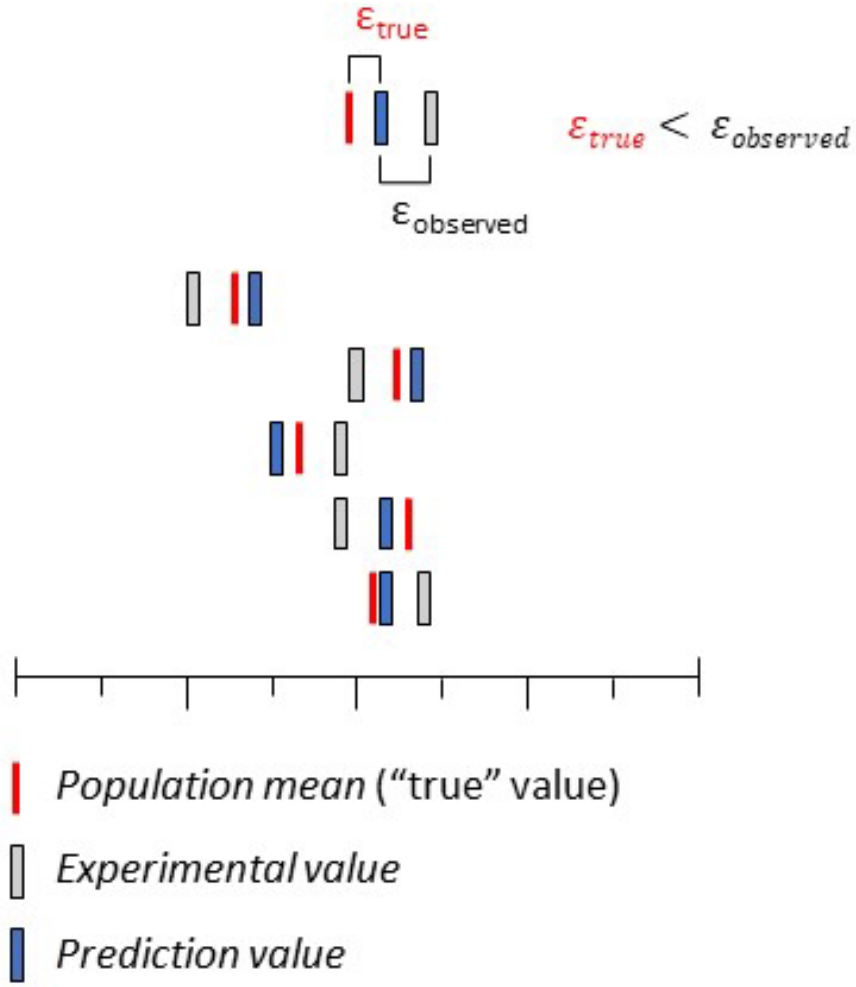
This work does not reflect EPA policy.

Evaluating QSAR Models



QSAR models attempt to predict the *population mean*

Evaluating QSAR Models



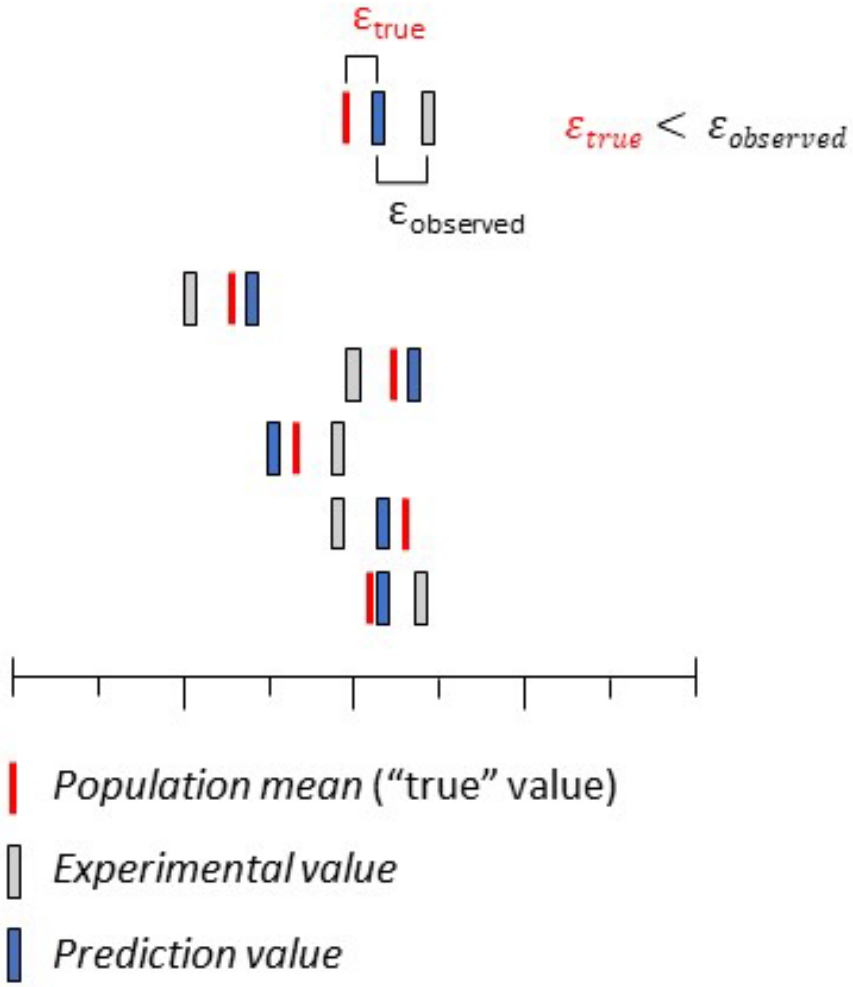
QSAR models attempt to predict the *population mean*

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

This evaluation is flawed however,
when the *experimental value* is not overlapping with the *population mean*;
this difference between them is ε_{true}

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 inevitably 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 inevitably trained on *experimental values* which do not represent *population means*?

Strategy

Take a QSAR dataset and:

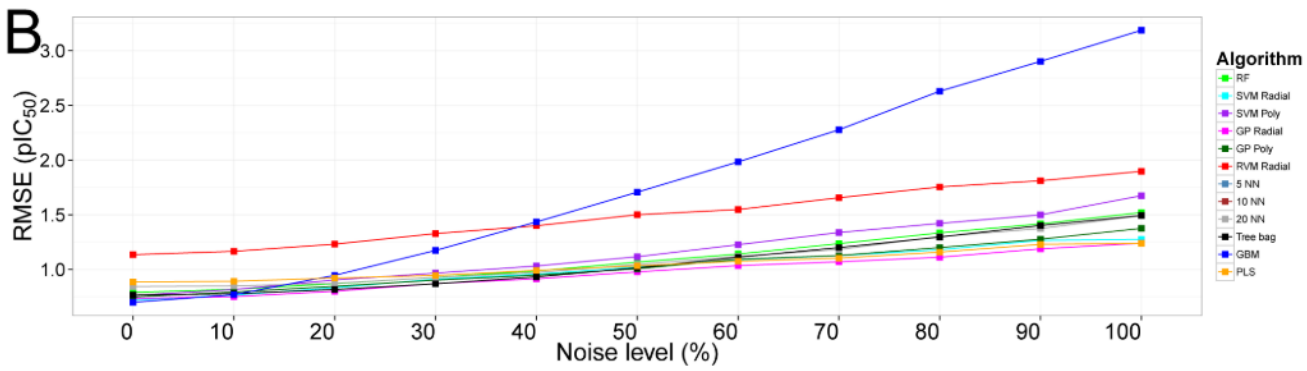
- Designate the original *experimental values* as "*population means*"
- Add simulated error to these values
- Predict the original values (*population means*)
- Predict the error laden values
- Compare metrics

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

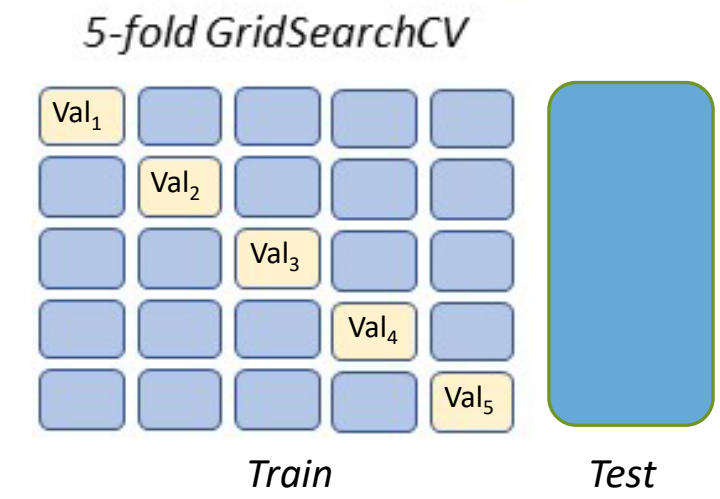
Error in QSAR

"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."

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



- *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
- So why is it often stated that a model's prediction accuracy is limited by experimental accuracy?

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*
- Predict the *true values*
- Predict the *error laden values*
- Compare model performance

Datasets

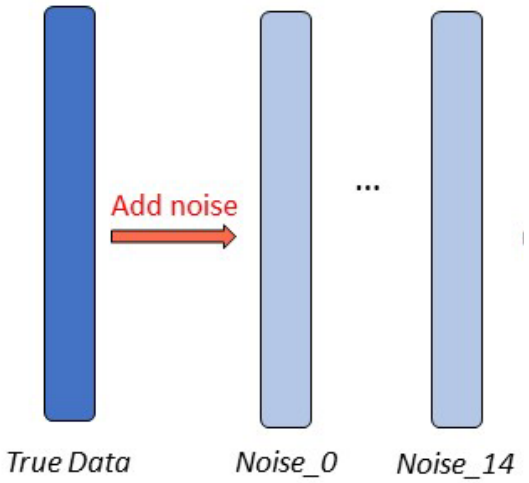
Dataset	Category	Number of Molecules[a]	Endpoint	Range
G298_atom	Quantum Mechanical	131,082	ΔG^{o}_{at} (kcal mol^{-1})	-2,417 – -288
Alpha	Quantum Mechanical	131,082	α (Bohr^{3})	9.0 – 27.8
Lip	Physiochemical	4,200	logD	-1.5 – 4.5
Solv	Physiochemical	642	ΔG^{o}_{hyd} (kcal mol^{-1})	-25.5 – 3.4
BACE	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.

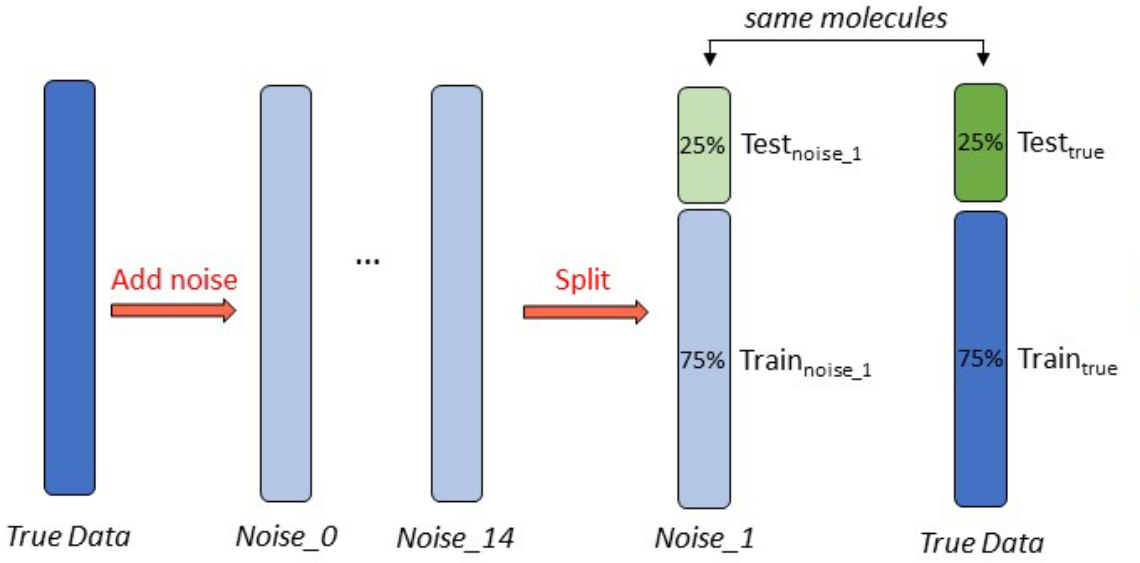
[b] Includes data exclusively from the ATG-PPre-cis assay

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

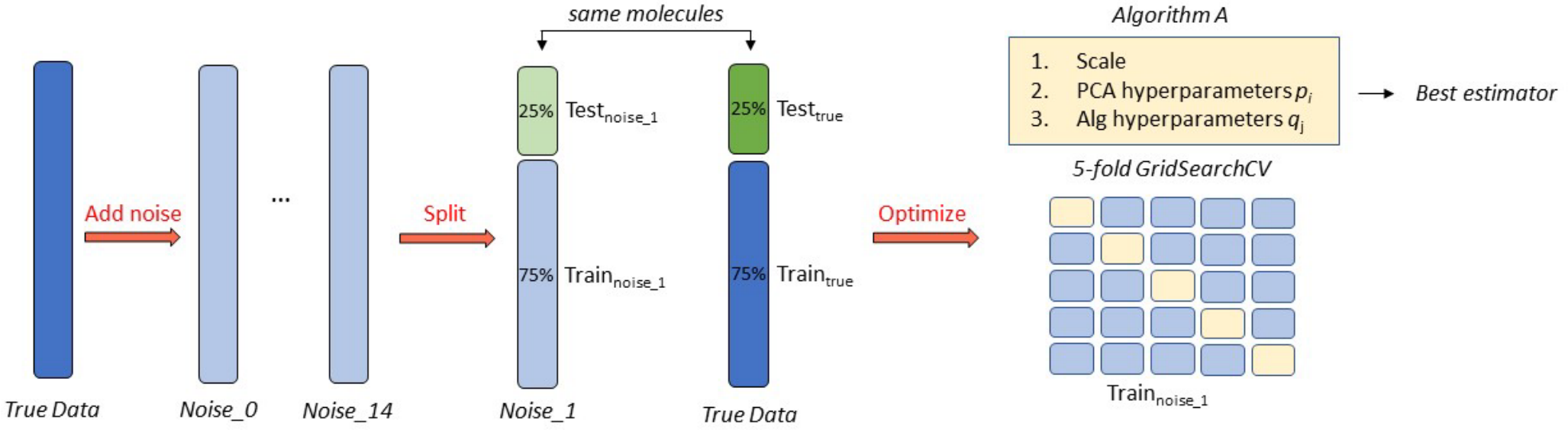
Modeling Workflow



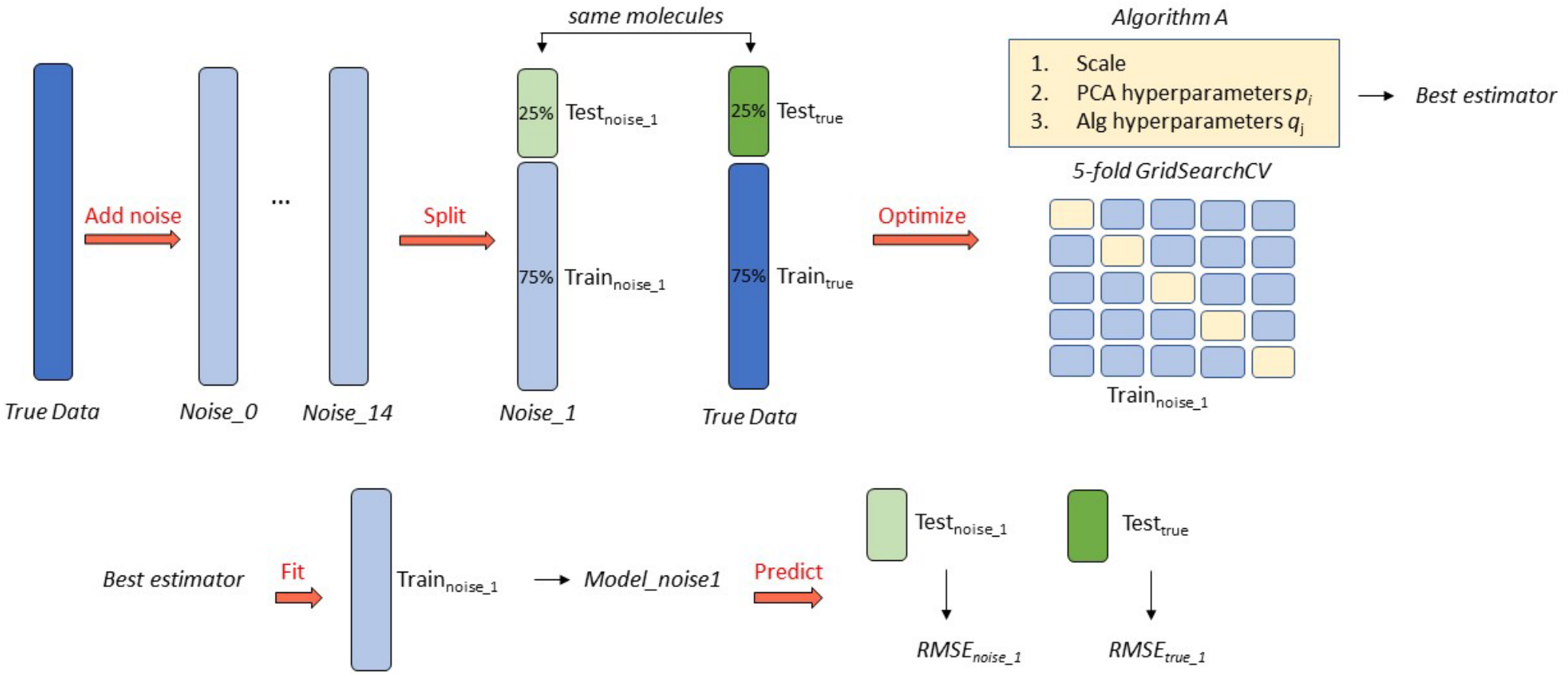
Modeling Workflow



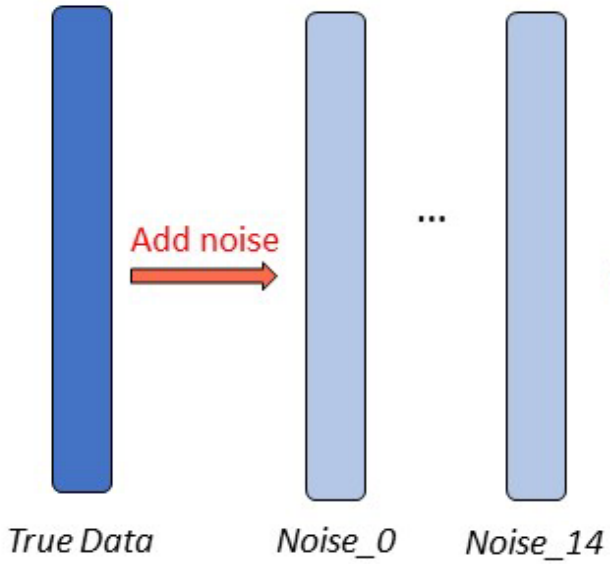
Modeling Workflow



Modeling Workflow



Simulating Random Error



$$Y_{noise_n,i} = Y + N(0, \sigma_{noise_n})$$

$$\sigma_{noise_n} = (Y_{max} - Y_{min}) * multiplier * n$$

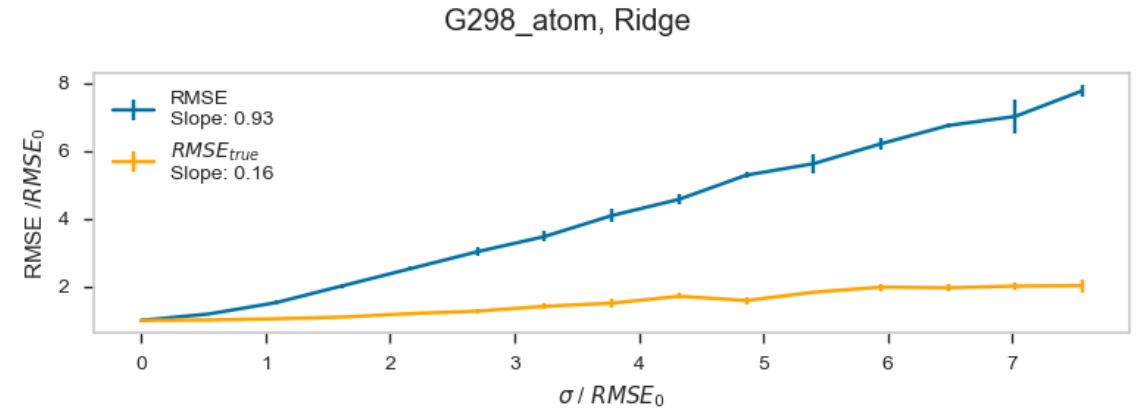
$$n \in (0, \ldots, 14)$$

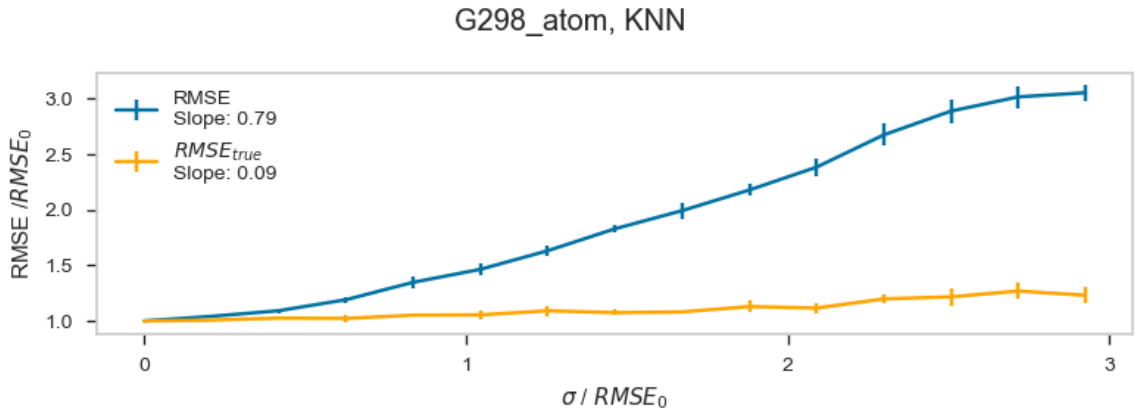
$$i \in (1, \ldots, 5)$$

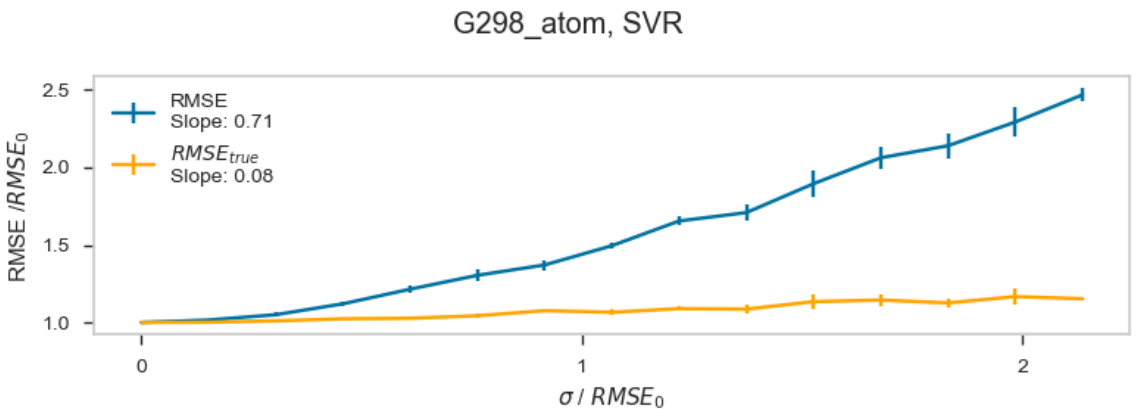
Algorithms and Hyperparameters

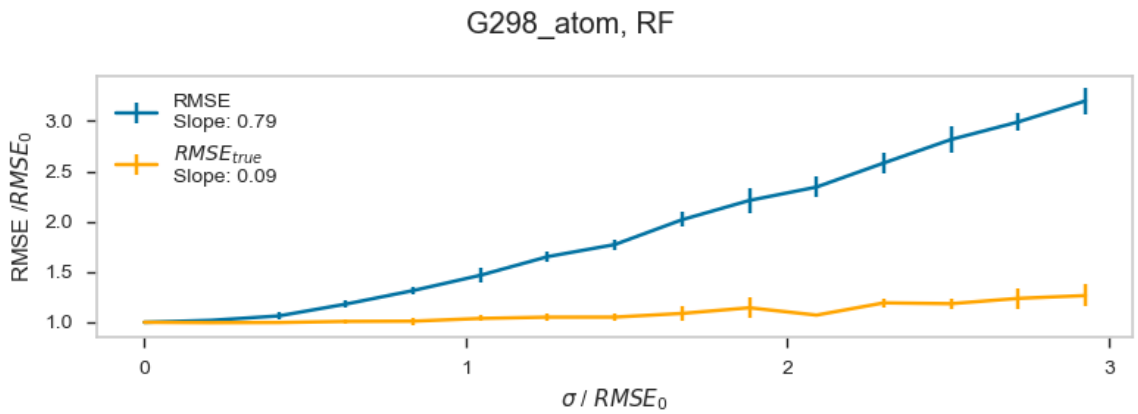
Algorithm	Hyperparameters Searched in Optimization
Ridge Regression (Ridge)	*PCA n components* $\in (1, 3, \ \dots\ , 59)$
	$\alpha \in (1, 2, 3, 4, 5, 10)$
k- Nearest Neighbors (kNN)	*PCA n components* $\in (1, 3, \ \dots\ , 59)$
	$k \in (1, 2, \dots, 20)$
Support Vector Regressor (SVR)	*PCA n components* $\in (1, 3, \ \dots\ , 59)$
	$C \in (0.01, 0.1, \ 1, 10)$
	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

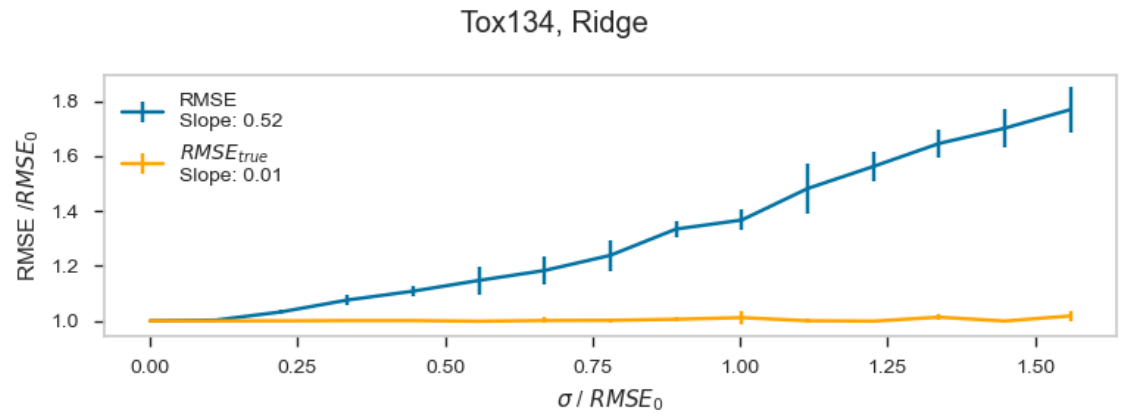


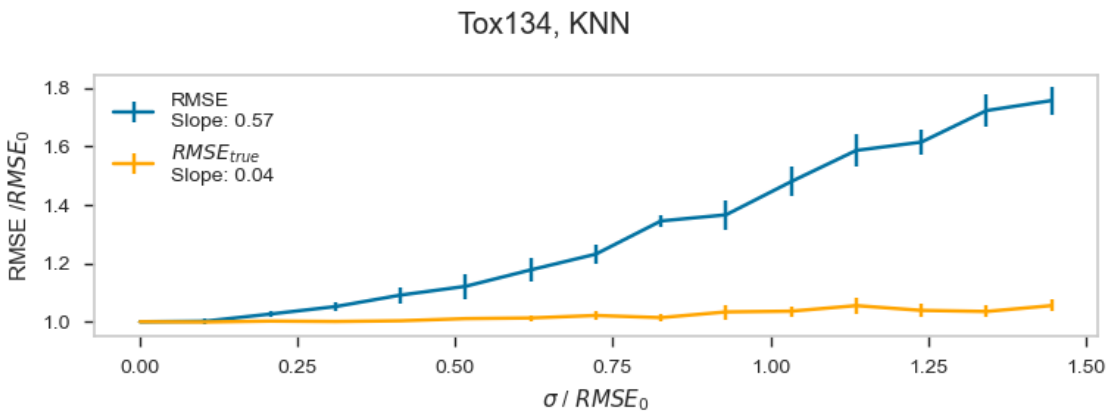


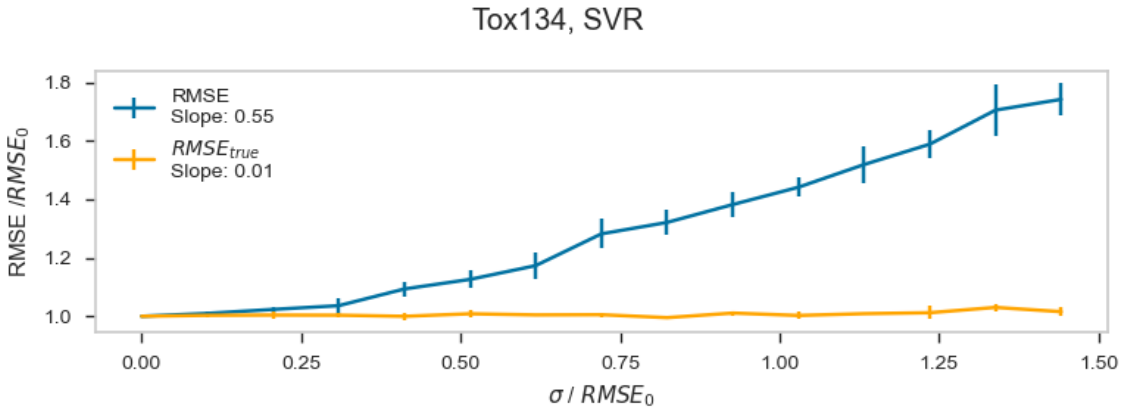


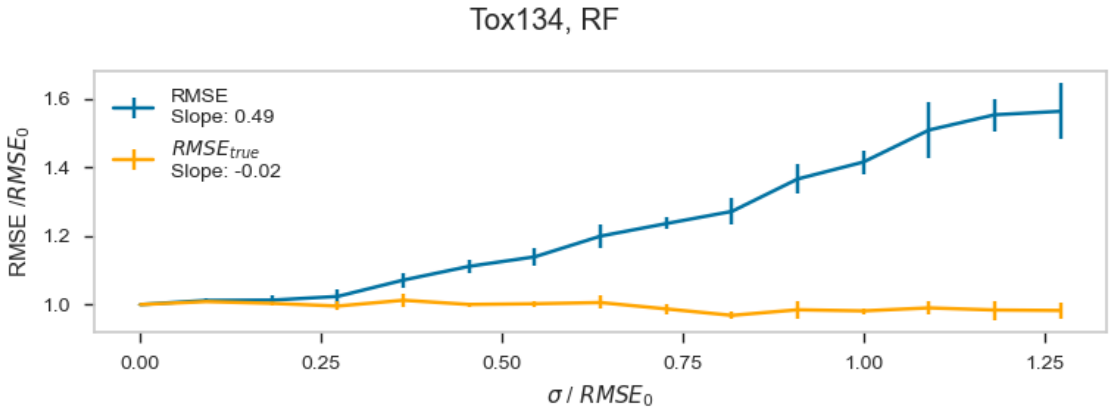


Tox134 Results









RMSE Slopes

Dataset	Slope	Ridge	kNN	SVR	RF	μ ± σ
G298_atom	m	0.93	0.79	0.71	0.79	0.81 ± 0.079
	m_{true}	0.16	0.09	0.08	0.09	0.11 ± 0.032
Alpha	m	1.0	0.83	0.87	0.89	0.90 ± 0.063
	m_{true}	0.14	0.10	0.12	0.12	0.12 ± 0.014
Lip	m	0.40	0.36	0.44	0.41	0.40 ± 0.029
	m_{true}	0.02	0.02	0.06	0.03	0.033 ± 0.016
Solv	m	0.75	0.81	0.89	0.72	0.79 ± 0.065
	m_{true}	0.13	0.27	0.27	0.12	0.20 ± 0.073

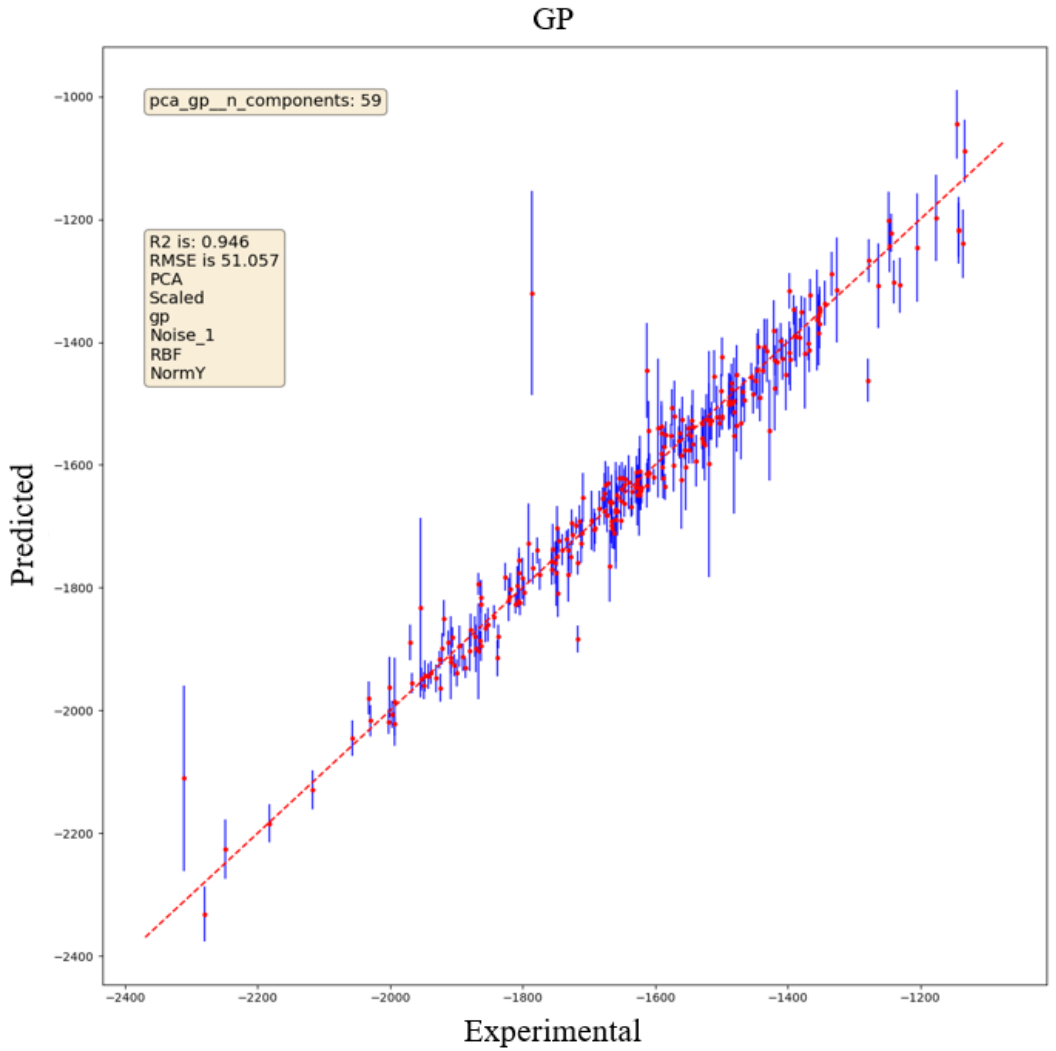
Dataset	Slope	Ridge	kNN	SVR	RF	μ ± σ
BACE	m	0.52	0.53	0.67	0.54	0.57 ± 0.061
	m_{true}	0.04	0.05	0.23	0.05	0.093 ± 0.079
Tox_102	m	0.44	0.49	0.44	0.43	0.45 ± 0.023
	m_{true}	0.01	0.05	0.002	0.01	0.018 ± 0.019
Tox_134	m	0.52	0.57	0.55	0.50	0.53 ± 0.027
	m_{true}	0.01	0.04	0.01	-0.02	0.01 ± 0.021
LD50	m	0.44	0.43	0.48	0.48	0.46 ± 0.023
	m_{true}	0.00	0.04	0.08	0.03	0.038 ± 0.029

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



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

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

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

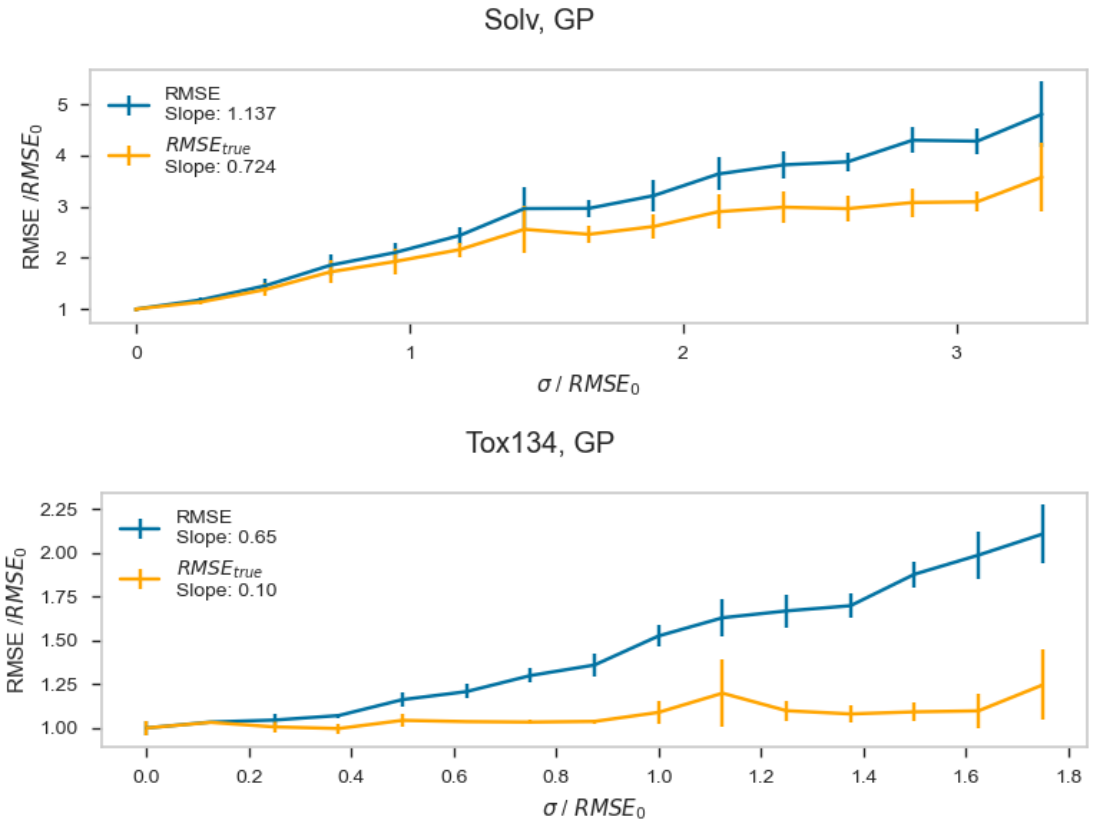
$$\sigma_{\hat{y}}\ 95\%\ CI = \frac{1.960}{\sqrt{n}}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\sigma_i - Mean\ \sigma_{\hat{y}}\right)^2\right]$$

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

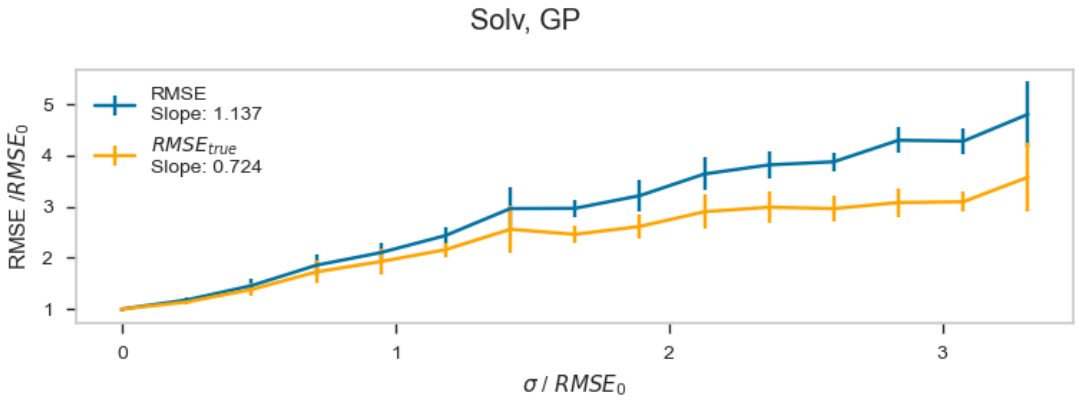
$$\sigma_y = \sigma_{y1}, \sigma_{y2}, \ldots, \sigma_{yn}$$

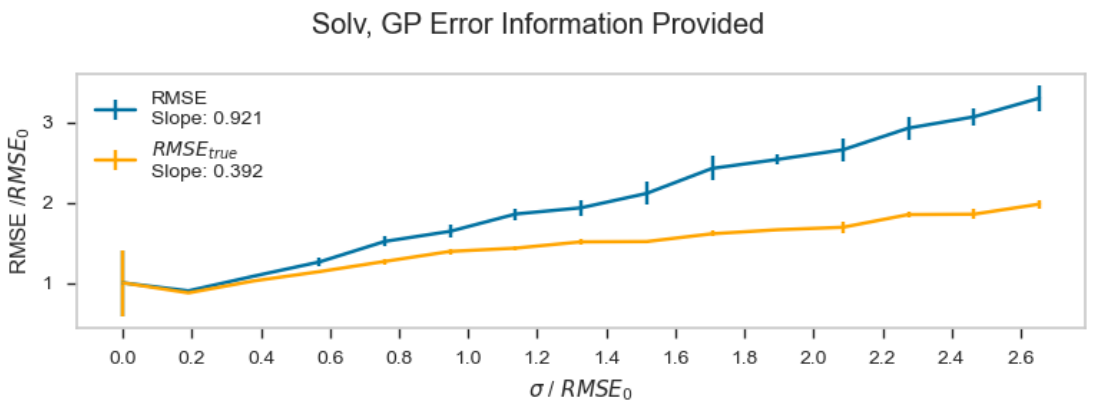
Information about experimental uncertainty

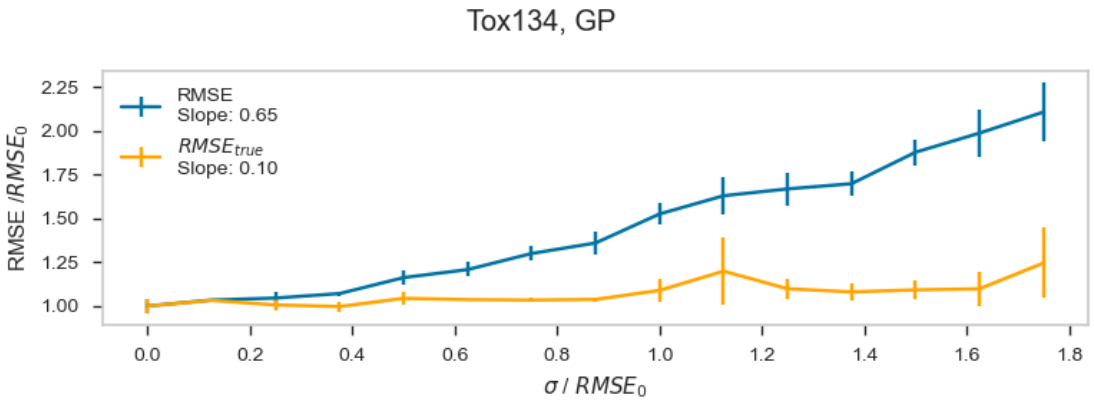
Gaussian Process (GP) Results

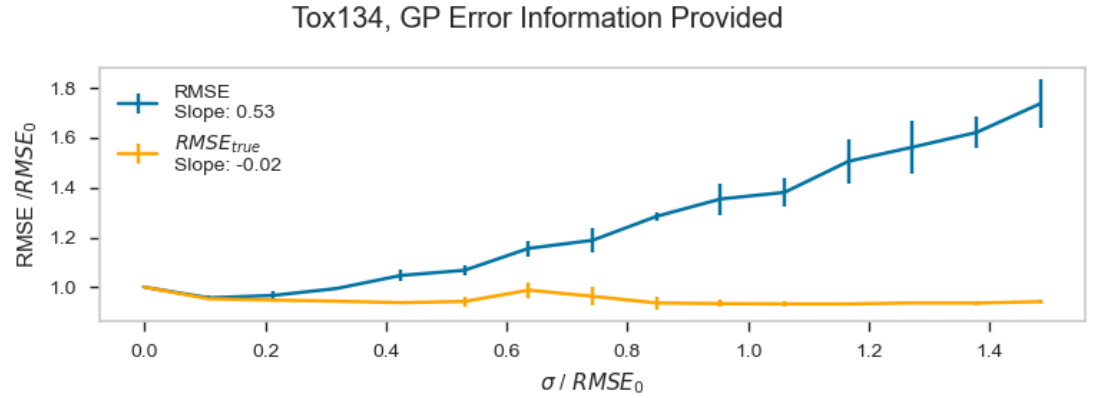


Gaussian Process (GP) Results









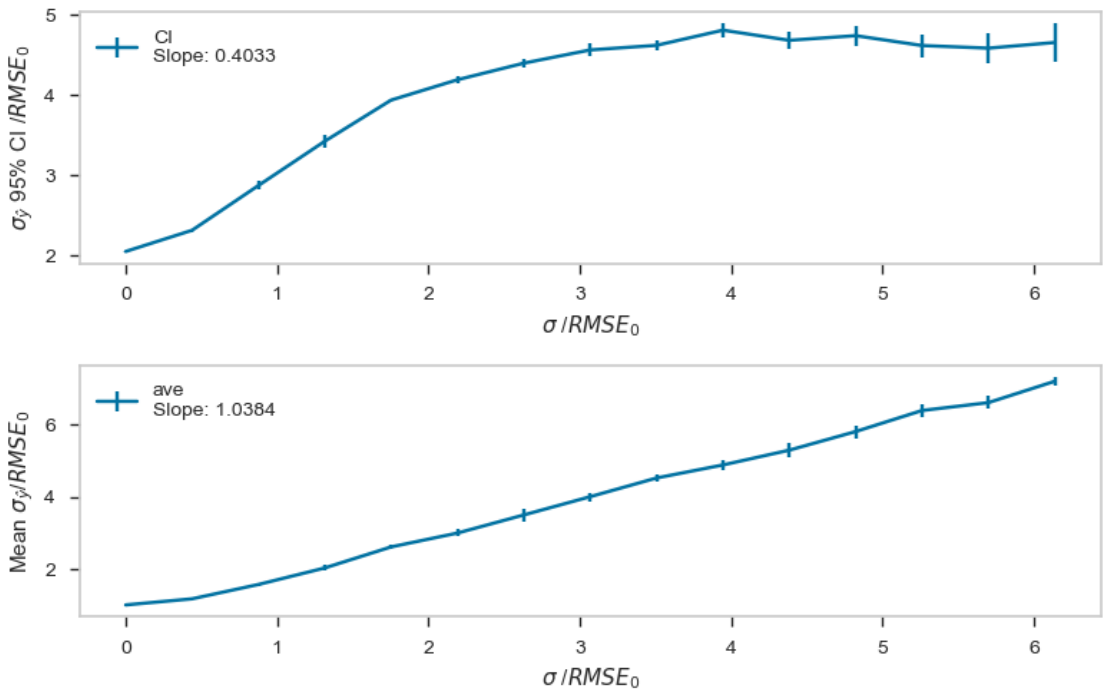
GP Slope ratios

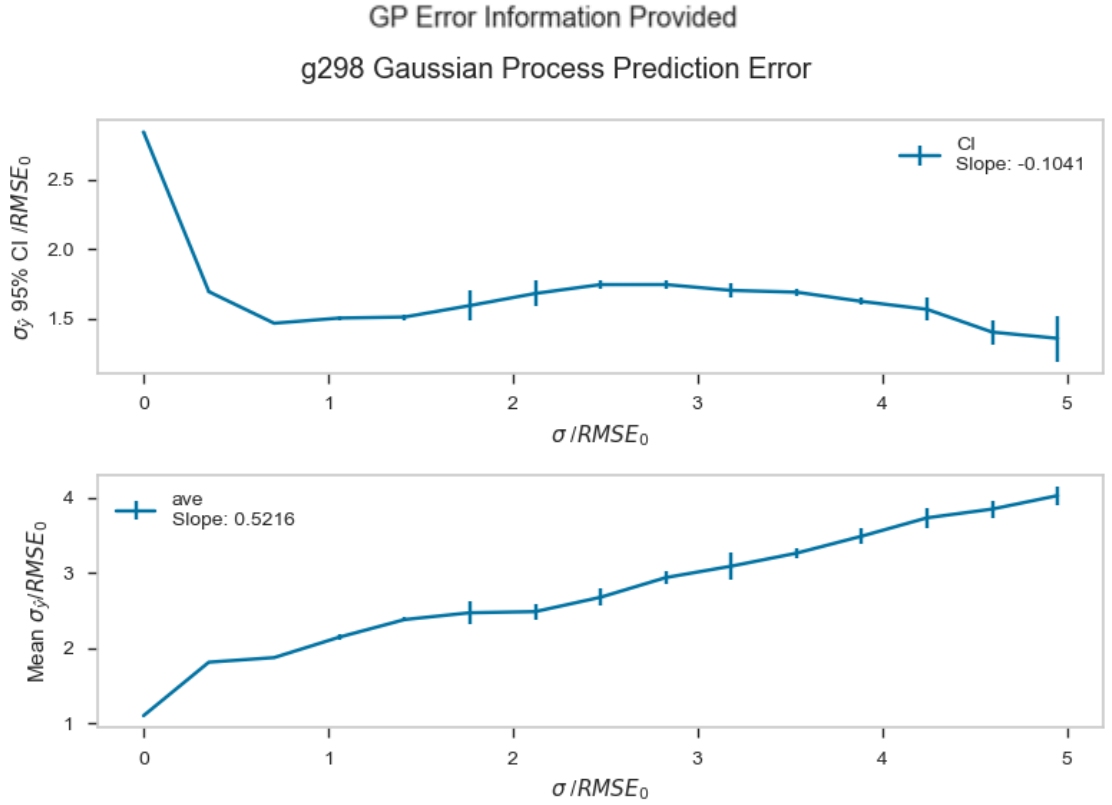
Dataset	*No* σ_y	*With* σ_y
G_298_atom	1.9	2.0
Alpha	1.8	9.4[a]
Solv	1.6	2.5[a]
BACE	3.8	7.8[a]
Tox_102	2.8	-[b]
Tox_134	7.0	-[b]
LD50	5.4	6.0
$\mu \pm \sigma$	3.5 ± 1.9	5.5 ± 2.9

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

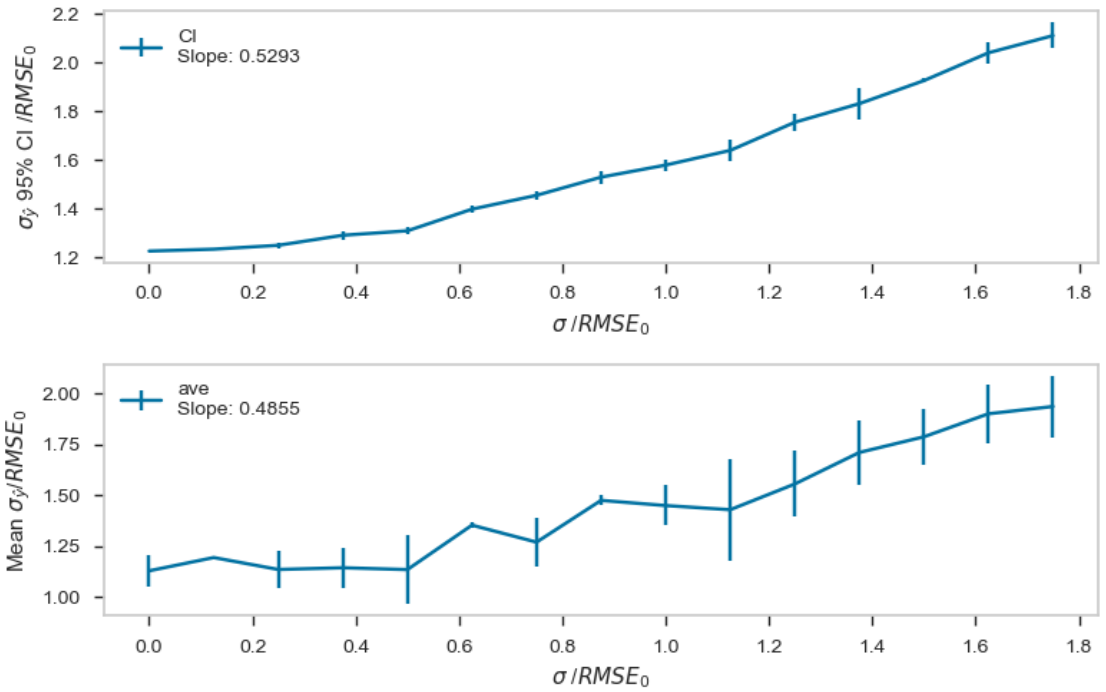
[b]The slope m_{true} was negative for these plots, so the slope ratio was not calculated.

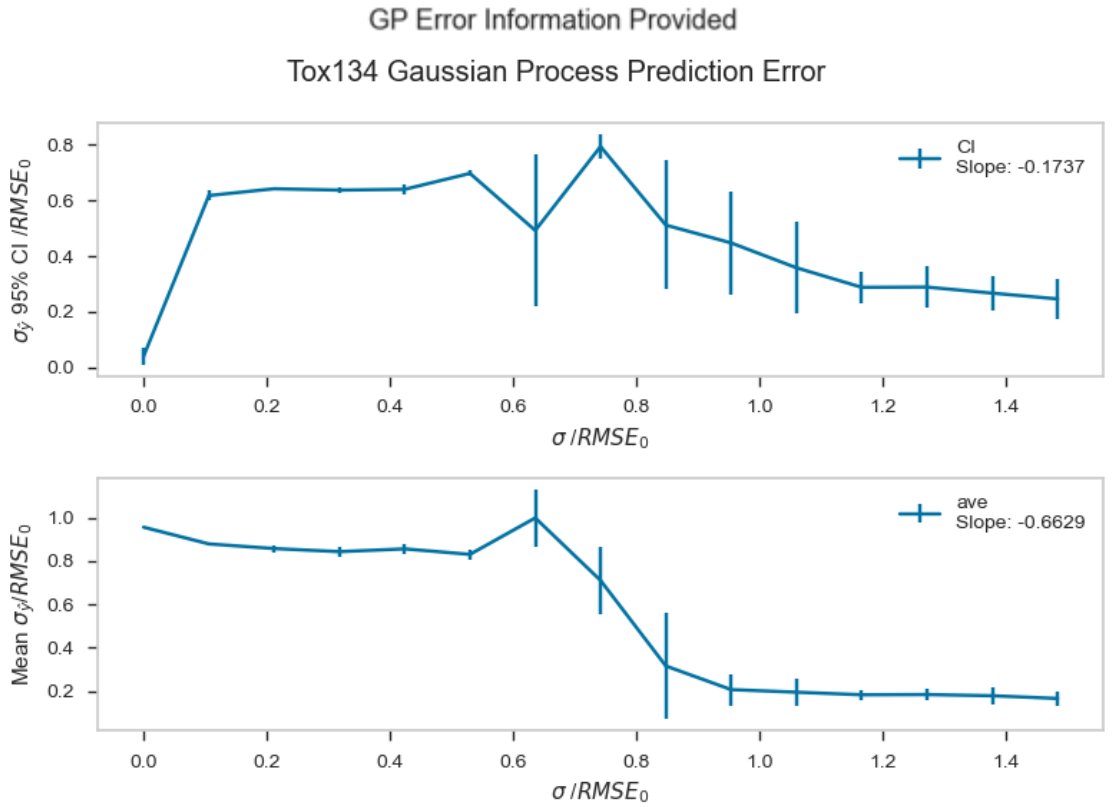
Gaussian Process (GP) Results





Gaussian Process (GP) Results





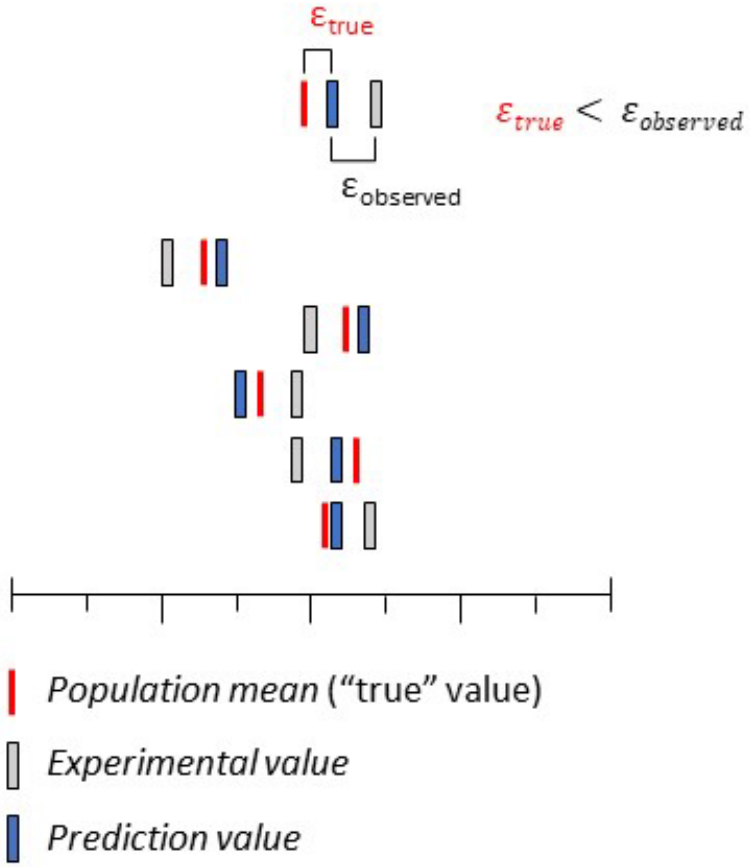
GP Prediction Uncertainties

Dataset	*No* σ_y *Mean* σ_y	*No* σ_y σ_y *95% CI*	*With* σ_y *Mean* σ_y	*With* σ_y σ_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

[a] The first point was omitted in these calculations because of a discontinuity in the line.

Conclusions

- QSAR models are built on data which typically do not approximate the population means of the measurements



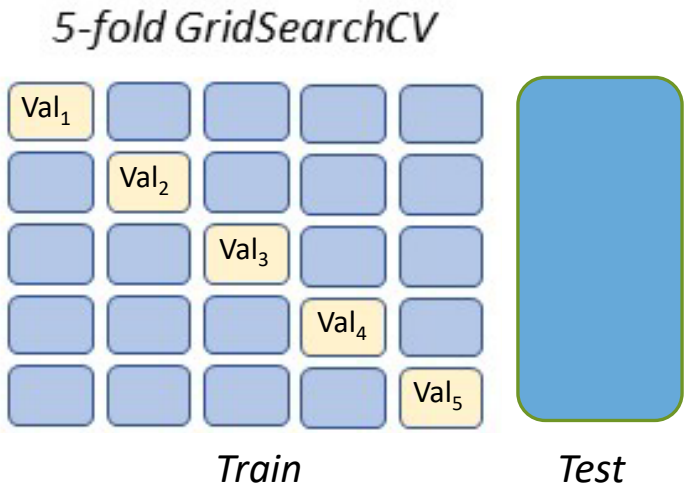
Wenlock et al. *J. Chem. Inf. Model.*, **2015**, 55, 125

Kalliokoski et al. *PLoS ONE*, **2013**, 8, e61007

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

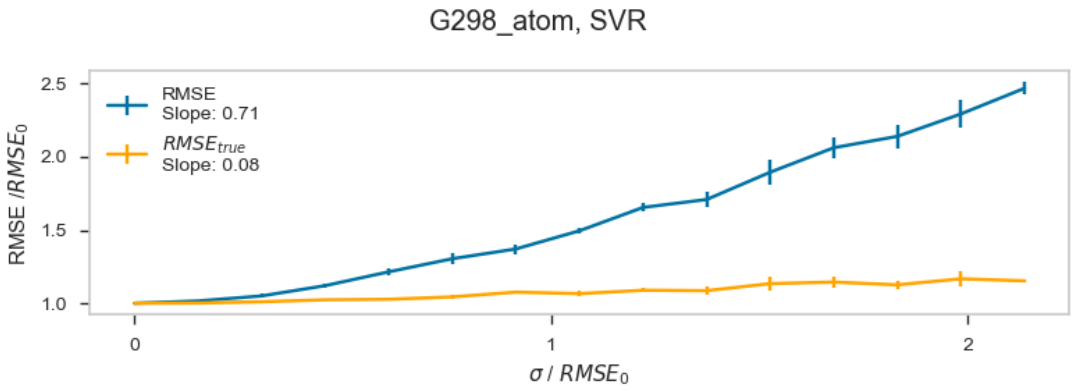
Conclusions

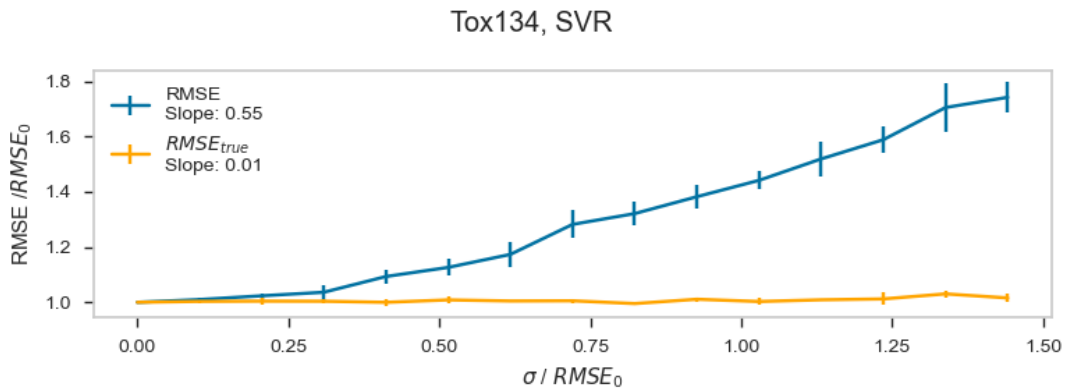
- QSAR models are evaluated on *Test* sets which have error



This has led to the assumption that a model's prediction uncertainty is limited by the experimental uncertainty in *Train*

Conclusions





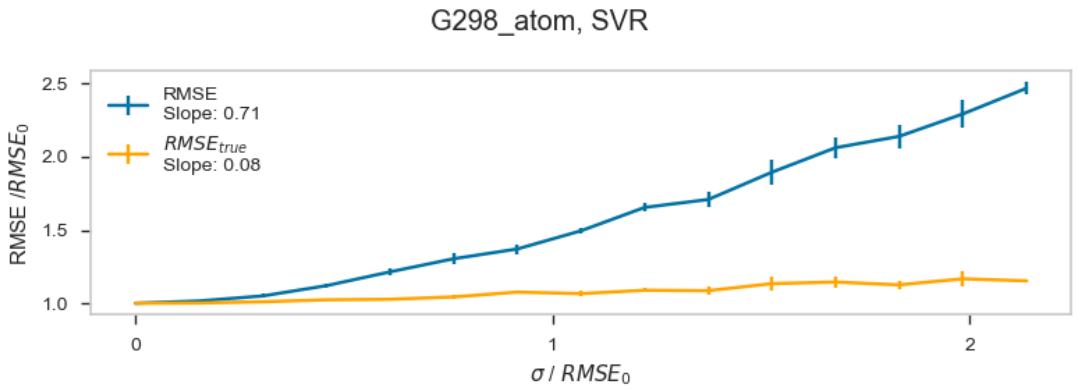
Methods

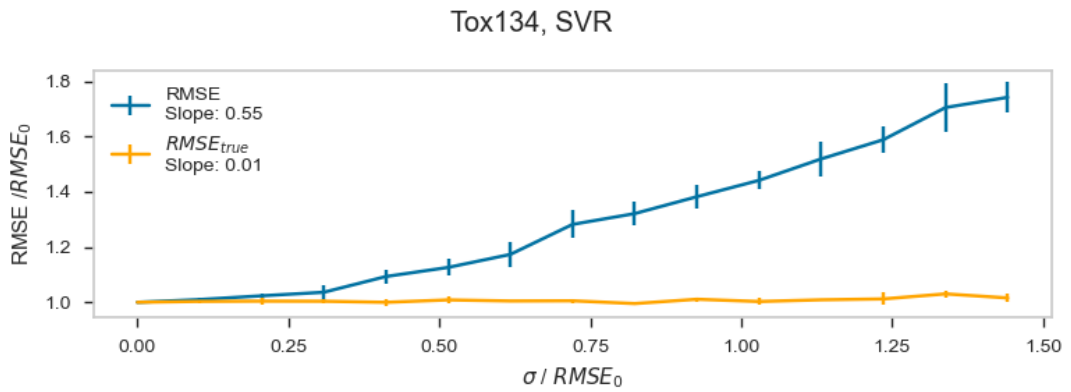
- Gaussian error was added to *8 representative QSAR datasets* and modeled using *5 algorithms*
  - The use of Gaussian distributed error represents an *ideal* but *realistic* simulation of real-world modeling

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 an algorithm which directly outputs prediction uncertainty such as Gaussian Process
  - Increasing the simulated error increases the prediction uncertainty
  - Providing information about error to the algorithm mitigates these trends

Conclusions





Implications

- QSAR models *can* predict population means accurately, even when trained on error laden values
- Evaluation of QSAR models on error laden test sets can give flawed interpretations of performance
  - A model may be predicting *population means* but this will be obscured by test set error

- Different models respond differently to error
  - $RMSE/RMSE_{true}$ is model dependent
  - *RMSE* is observed
  - *RMSEtrue* is unknown
- Determining relative performance between two different models could be tenuous and potentially misleading

Future Work

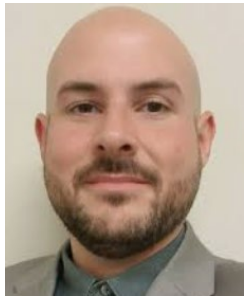
- Evaluation of new algorithms and new models will be similarly limited by knowledge of the uncertainty in validation and test sets
- New methods of inferring uncertainty in datasets and new evaluation methodologies which utilize knowledge of uncertainty are needed to give more reliable comparisons of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

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