New approach to regression uncertainty analysis and applications to drug design

Marvin Waldman and Robert D. Clark

Simulations Plus, Inc.

Lancaster CA, USA

SimulationsPlus

SCIENCE + SOFTWARE = SUCCESS

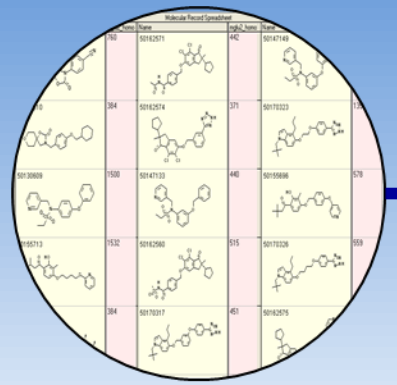
Overview

- Motivation
- Approaches/Methodology
- Application/Results
- Summary/Conclusions

Motivation

- When can you trust a decision/prediction from a machine learning model?
  - Many examples of machine learning/AI failures (just Google “recent AI failures”)
- What is the “expected” accuracy of a quantitative prediction?
  - Drug candidate with predicted low solubility
    - “Distrust” the model – expected accuracy is poor – large prediction uncertainty
      - Synthesize anyway and measure the solubility
    - “Trust” the model – expected accuracy is good – small prediction uncertainty
      - Move on to another compound – don’t bother to synthesize

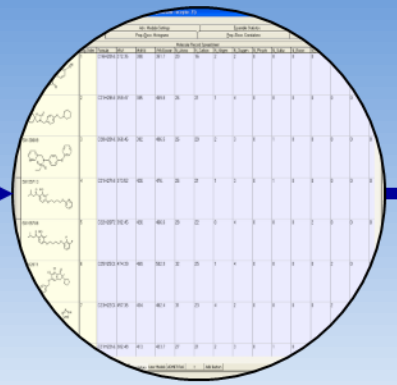
Model Building - Overview



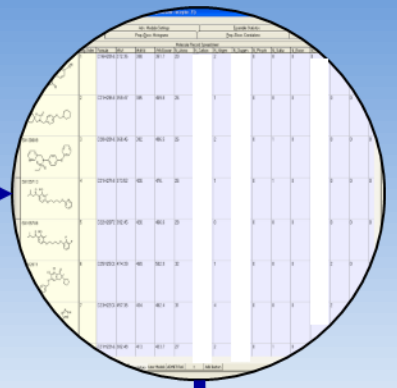
Structures and experimental data

Compute >300 Descriptors

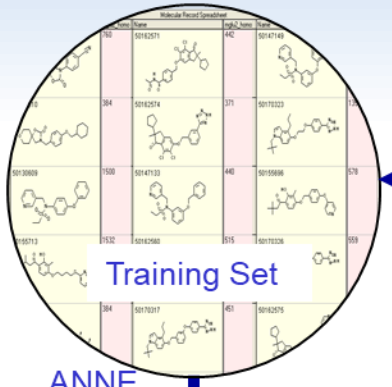
Constitutional
Topological
Hbonding
Ionization
Electrotopological
Charge/reactivity



Remove low variance and correlated descriptors

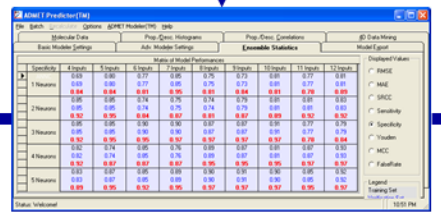


Test Set Selection
1) Kohonen map
2) Every n-th
3) Random
4) K-means
5) Manual



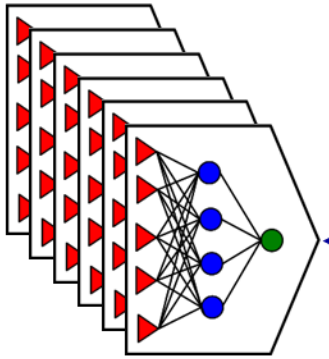
- No. of neurons and descriptors
  - Create models with different architectures
- Sensitivity analysis
  - Which descriptors create the best model

ANNE Training



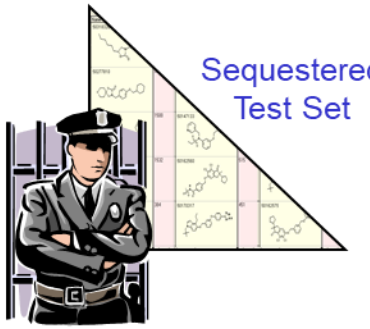
Grid of ANNE Models
(X descriptors by Y neurons)

Select best model



Ensemble of 33 ANNs

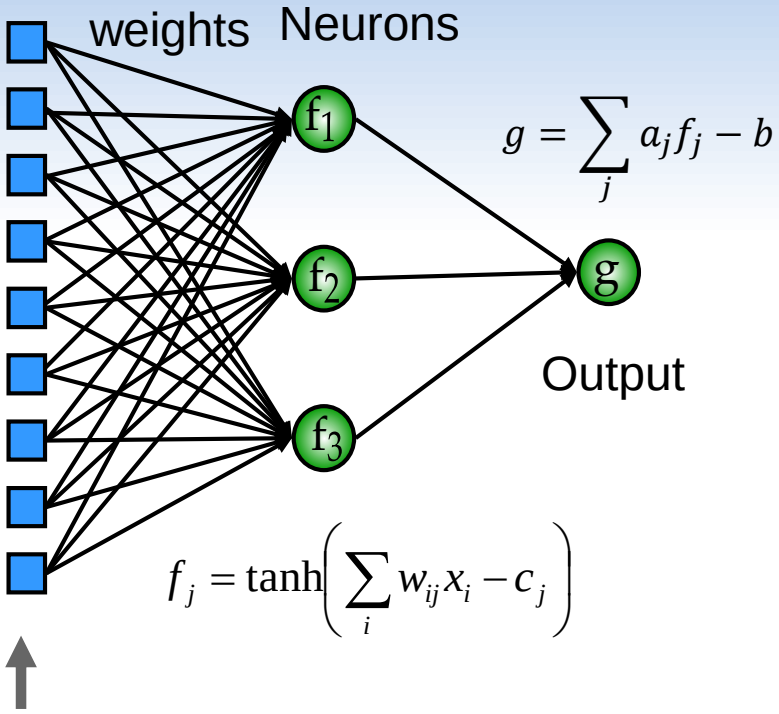
Apply model to test set



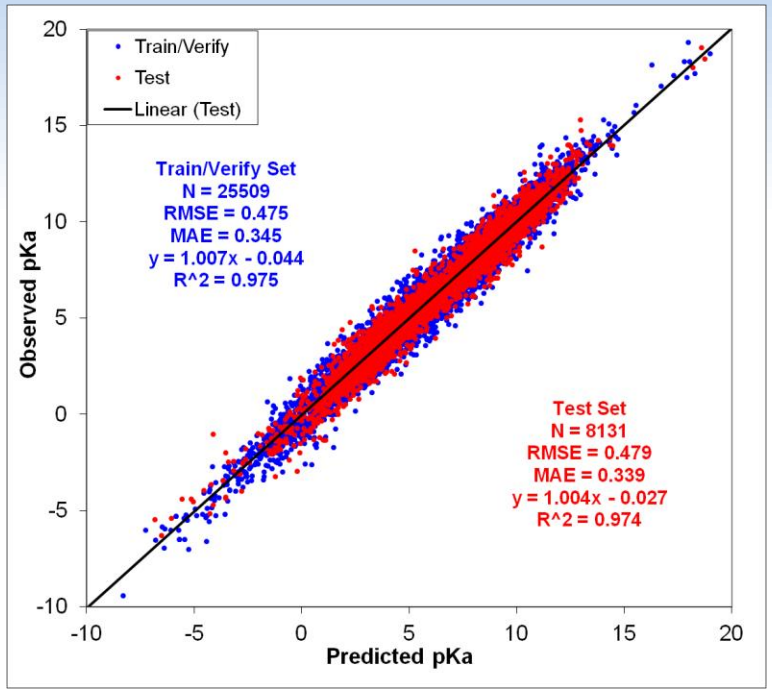
Sequestered Test Set

SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

Artificial Neural Network (ANN) Architecture



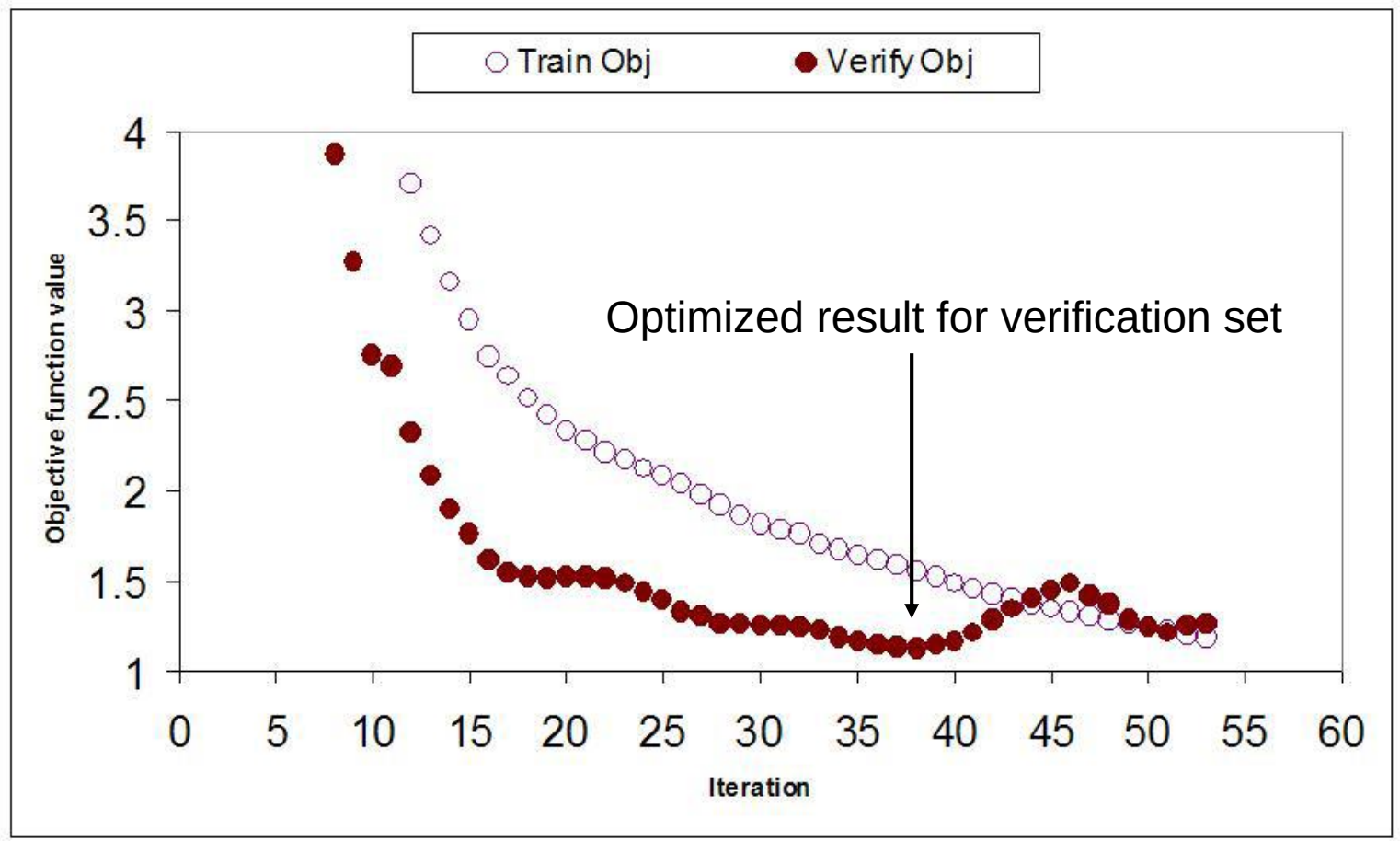
Molecular and atomic descriptors



SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

Avoiding Overtraining: Early Stopping

1. Split training set into training and verification sets
2. Optimize network weights to improve training set performance
3. Monitor performance of verification set – determines stopping point



SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

Artificial Neural Network Ensembles (ANNE)

- Repeat training/verify random split 165 times – select best 33 networks
- Model Prediction is average of 33 network predictions ($\hat{y}_i$)

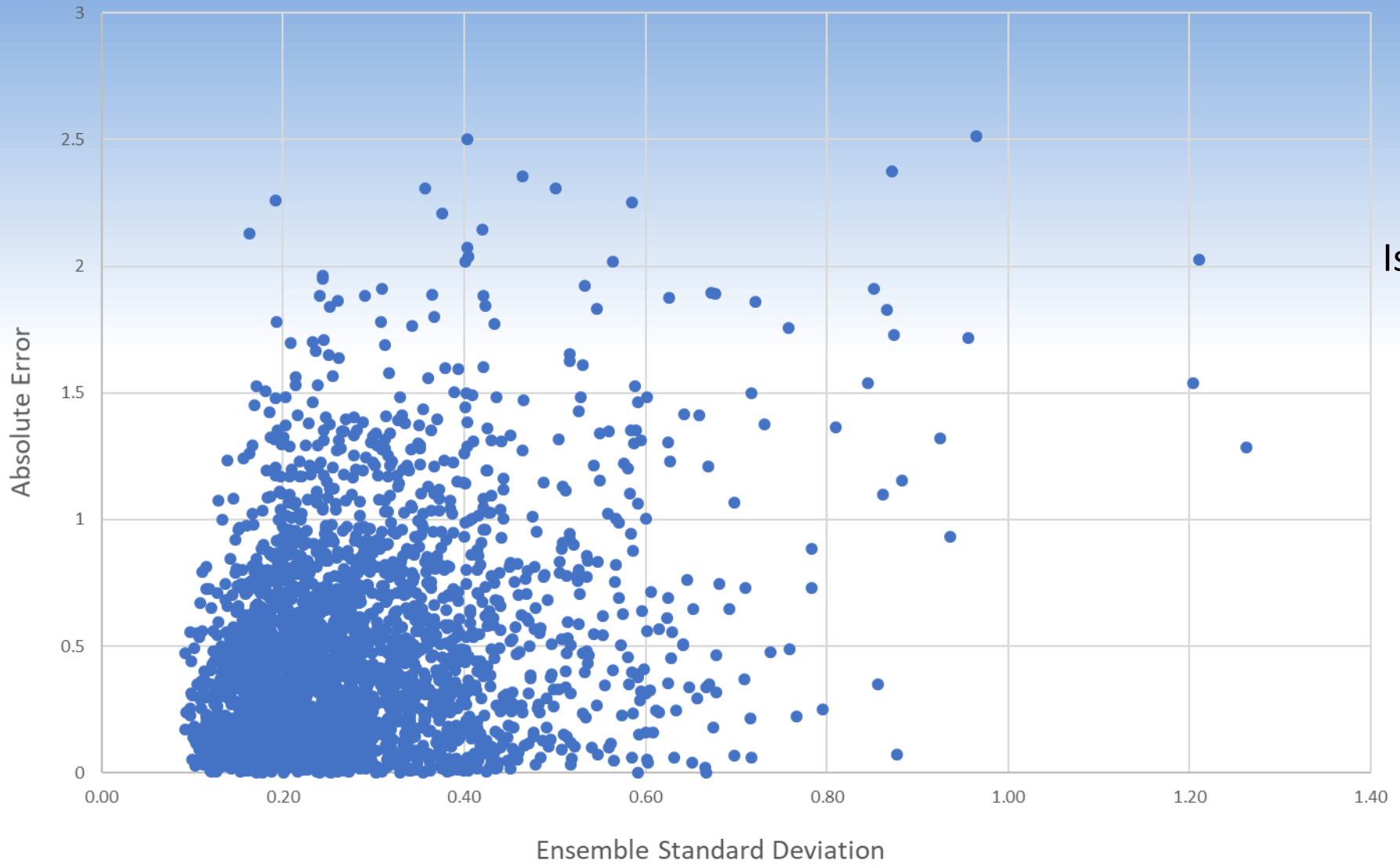
$$\bar{\hat{y}} = {\sum_i \hat{y}_i}/{N}$$

- What about the variance of the individual network predictions?

$$\sigma^2 = \frac{\sum_i \left(\hat{y}_i - \bar{\hat{y}}\right)^2}{N-1}$$

  - Can this be used to assess uncertainty in the prediction?
- Previously, we showed how to estimate uncertainty in classification prediction from the degree of disagreement among the individual network predictions
  - Clark et al., J. Chem. Informatics, “Using beta binomials to estimate classification uncertainty for ensemble models” **6** 34 (2014)
- What about regression models (continuous output)?

At first glance …



Is there signal in this noise?

SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

Earlier Work

Beck et al., *J. Chem. Inf. Comput. Sci., Vol. 40, No. 4, 2000* **1049**

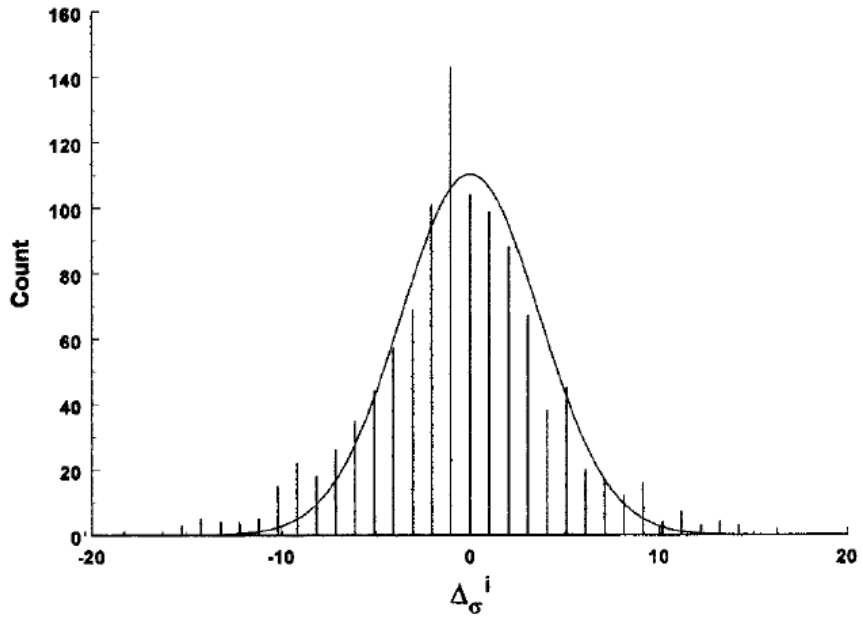


Figure 5. The distribution of Δ_σ^i-values for the 11-net logP model. The solid line shows the best fit Gaussian.

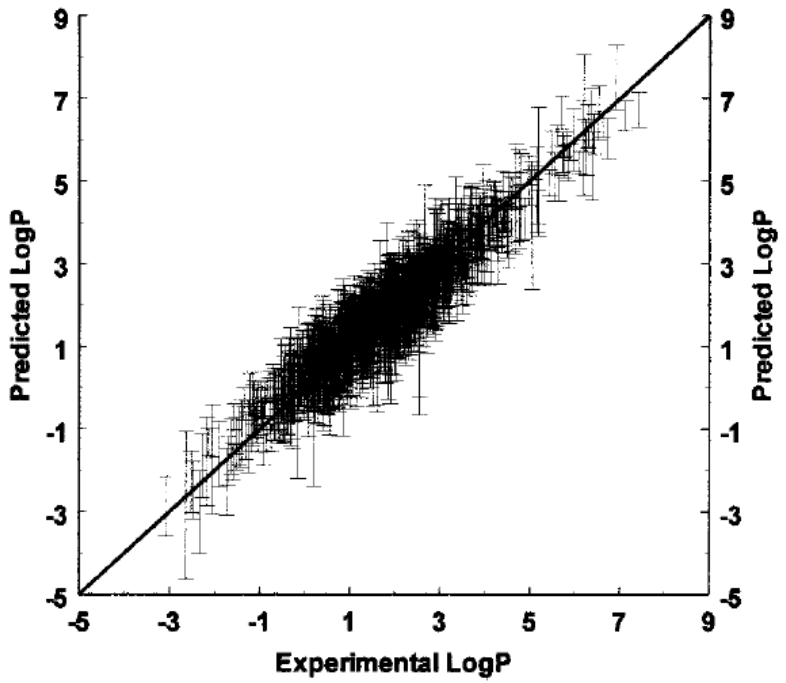


Figure 7. Results for the logP dataset ($N = 1085$) with the mean values and error estimations given in eq 2.

$$\Delta_\sigma^i = \frac{(P_{calc}^i - P_{obs}^i)}{\sigma^i}$$

$$P_i = \overline{P}_i \pm \overline{\Delta}\sigma_i$$

Uncertainty is assumed to be proportional to standard deviation of ensemble prediction

Observed MAE (Mean Absolute Error) = 0.38

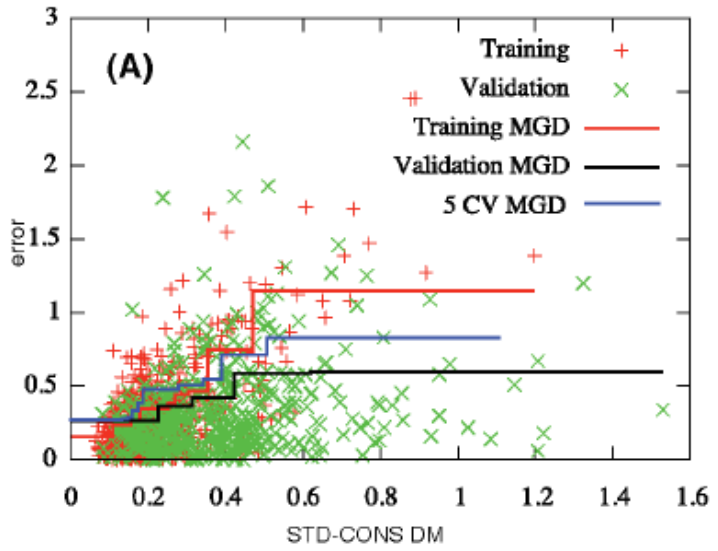
Calculated MAU (Mean Absolute Uncertainty) = 0.50

SimulationsPlus

SCIENCE + SOFTWARE = SUCCESS

Earlier Work (Part II)

Tetko et al.



Errors are binned with respect to Ensemble Std. Dev. and a Gaussian is fitted to each bin – width of Gaussian is uncertainty estimate for that bin

Table 3. Performances of MGDs on the Training and on the Joint Validation Sets

	average rank		
DM[a]	LOO	5-CV	valid.
STD-CONS	1	1.8	1.1
STD-ASNN	2	1.2	2.5
STD-kNN-DR	6.6	4.3	4.1
STD-kNN-MZ	9.2	8.3	5.3
EUCLID-kNN-DR	7.1	4.9	5.4
LEVERAGE-PLS	8.4	5	6.3
EUCLID-kNN-MZ	7.5	7.1	6.4
TANIMOTO-kNN-FR	7	6.1	6.8
TANIMOTO-MLR-FR	8.3	8.3	9
CORREL-ASNN	10.7	10.8	9.4
LEVERAGE-OLS-DR	12.3	12.6	11.1
EUCLID-MLR-FR	7	9.3	11.5
PLSEU-PLS	11.1	11.8	11.5
EUCLID-kNN-FR	12.1	13.3	12.1

Ensemble Std. Dev. performed best as a metric of uncertainty – ability to discriminate between small and large errors.

Earlier Work (Part III)

- Conformal Prediction

Cortes-Ciriano and Bender, JCIM, 59, 1269 (2019) – Deep Confidence

$$Confidence\ region = \hat{y}_j \pm (\alpha_{CL})e^{\sigma_j}$$

"σ_j is the standard deviation of predicted activities across the ensemble."

- What is the basis of the exponential dependence on standard deviation?
  - Papadopoulos et al., J. Artificial Intell. Res. 40 815 (2011)
  - Conformal prediction using exponential of scaled std. dev. of k-NN predictions
  - Justification:
    - "The exponential function in definition (25) was chosen because it has a minimum value of 1, since σ will always be positive, and grows quickly as σ increases. As a result, this measure is more sensitive to changes when σ is big, which indicates that an example is unusually far from the training examples."

(Temporarily) return to the binning approach …

$$Global\ RMSE = \sqrt{\frac{SSE}{N}}$$

Where

$$SSE = \sum_i (\hat{y}_i - y_i)^2$$

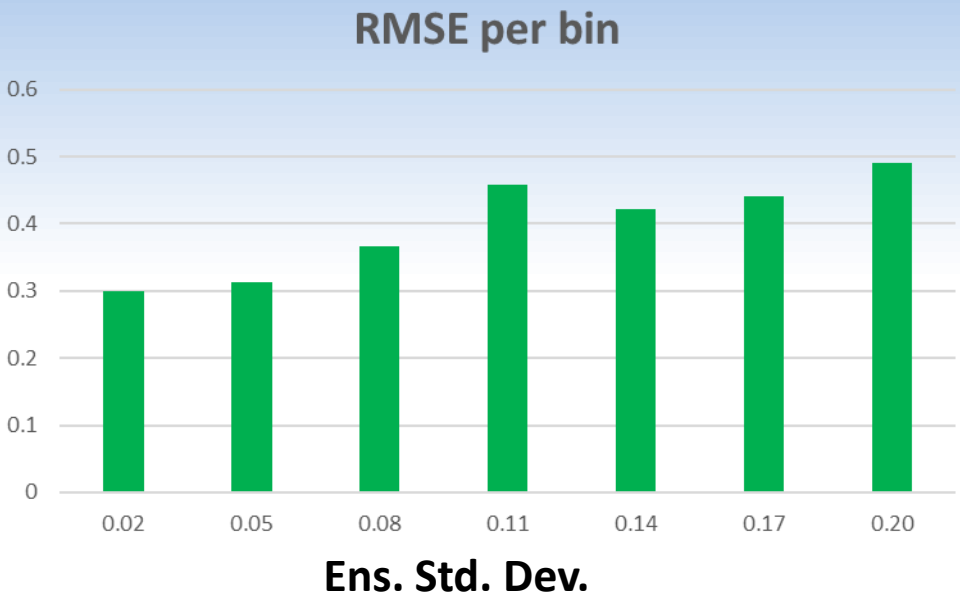
Form bins over ensemble std. dev.

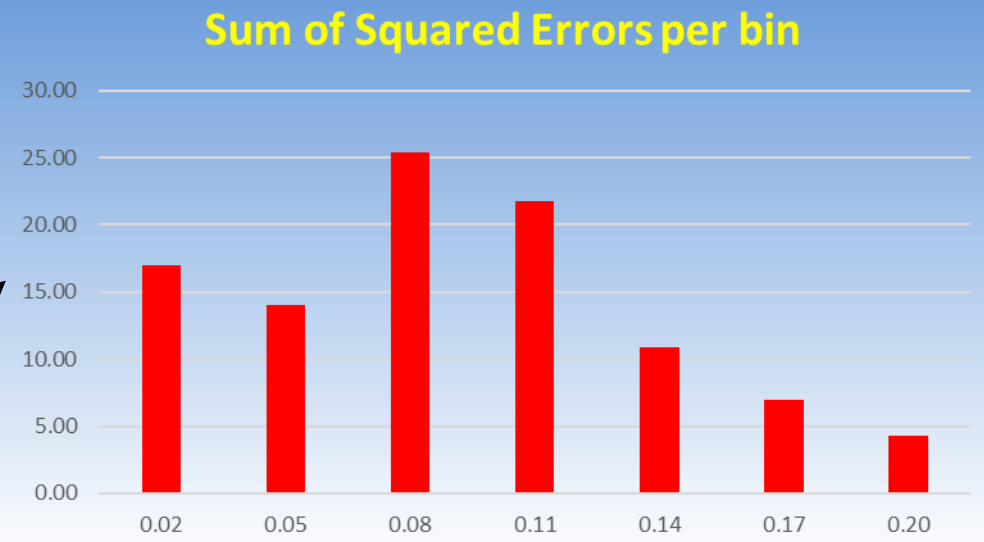
$$Local\ RMSE = \sqrt{\frac{SSE_j}{N_j}}$$

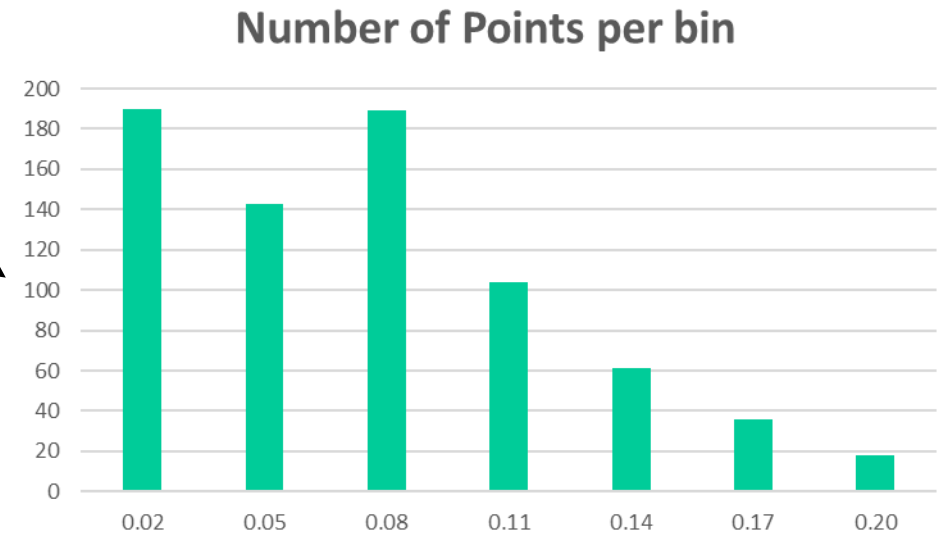
← Sum of squared errors in each bin

← Number of points in each bin

Graphically …







SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

Removing the bins …

$$f_j = \frac{N_j}{N} \qquad g_j = \frac{SSE_j}{SSE}$$

$$RMSE_j = \sqrt{\frac{SSE * g_j}{N * f_j}} = RMSE\sqrt{\frac{g_j}{f_j}}$$

$$u(s) \equiv RMSE(s) = RMSE\sqrt{\frac{g(s)}{f(s)}}$$

Uncertainty is defined as the "local" RMSE

s is the ensemble std. dev. and g(s) and f(s) are probability distributions of the fractional squared error and number of points with respect to s.

Formal derivation (for the mathematically inclined)

Bayes' Theorem

$$p(\varepsilon, s) = q(\varepsilon|s) f(s)$$ ← Probability of s

Joint probability of error ε and Ens. std. dev. s

Conditional probability of ε given s

$$\Phi(s) \equiv \int_{-\infty}^{\infty} d\varepsilon \int_{0}^{s} \varepsilon^2 p(\varepsilon, \chi) d\chi = \int_{0}^{s} f(\chi) d\chi \int_{-\infty}^{\infty} \varepsilon^2 q(\varepsilon|\chi) d\varepsilon$$

$$\Phi(s) = \int_{0}^{s} \sigma^2(\chi) f(\chi) d\chi \qquad Note: \ \Phi(\infty) = RMSE^2$$

$$\phi(s) \equiv \Phi'(s) = \sigma^2(s) f(s)$$ Fundamental Theorem of Calculus

$$\sigma^2(s) = \frac{\phi(s)}{f(s)} = RMSE^2 \frac{g(s)}{f(s)}$$

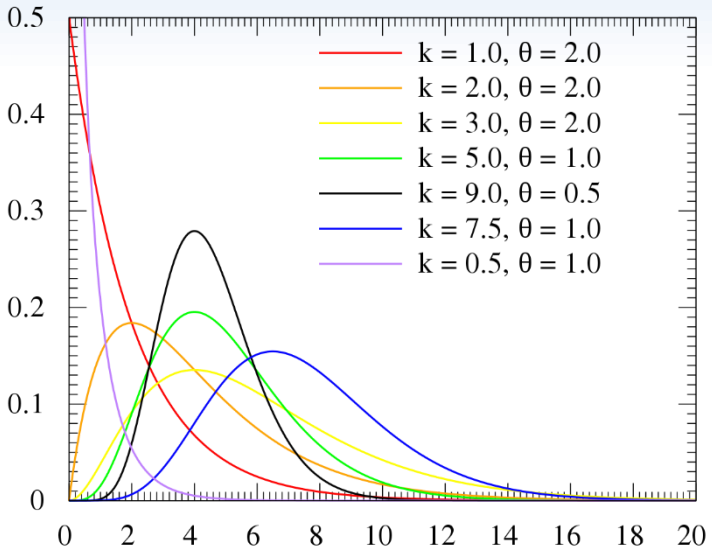
$$u(s) \equiv \sqrt{\sigma^2(s)} = RMSE \sqrt{\frac{g(s)}{f(s)}}$$

Choice of Distribution

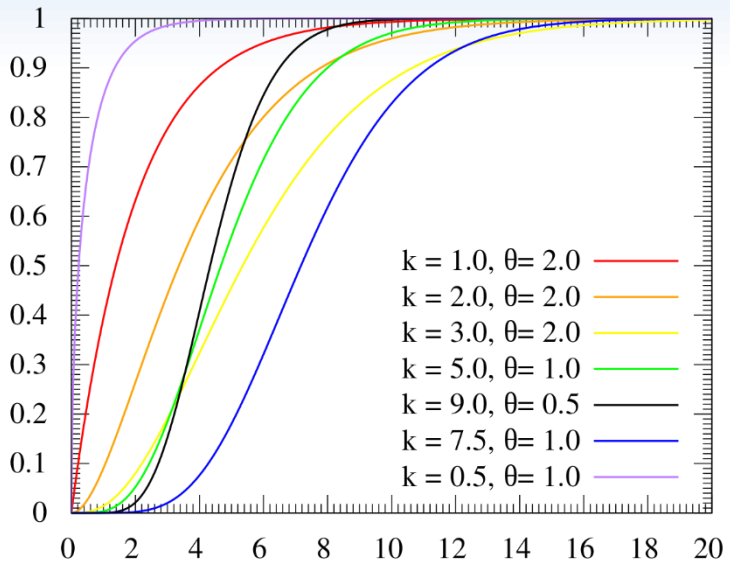
Gamma Distribution – a generalized Chi-squared

$$p(x;\alpha,\beta) = \frac{(\beta x)^{\alpha} e^{-\beta x}}{x\Gamma(\alpha)} \qquad \text{for } x > 0 \text{ and } \alpha, \beta > 0$$

Let $x = s - s_0$: s is Ens. std. dev. and s_0 is an added shift parameter



density



cumulative

https://en.wikipedia.org/wiki/Gamma_distribution

SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

Making it physically “reasonable”

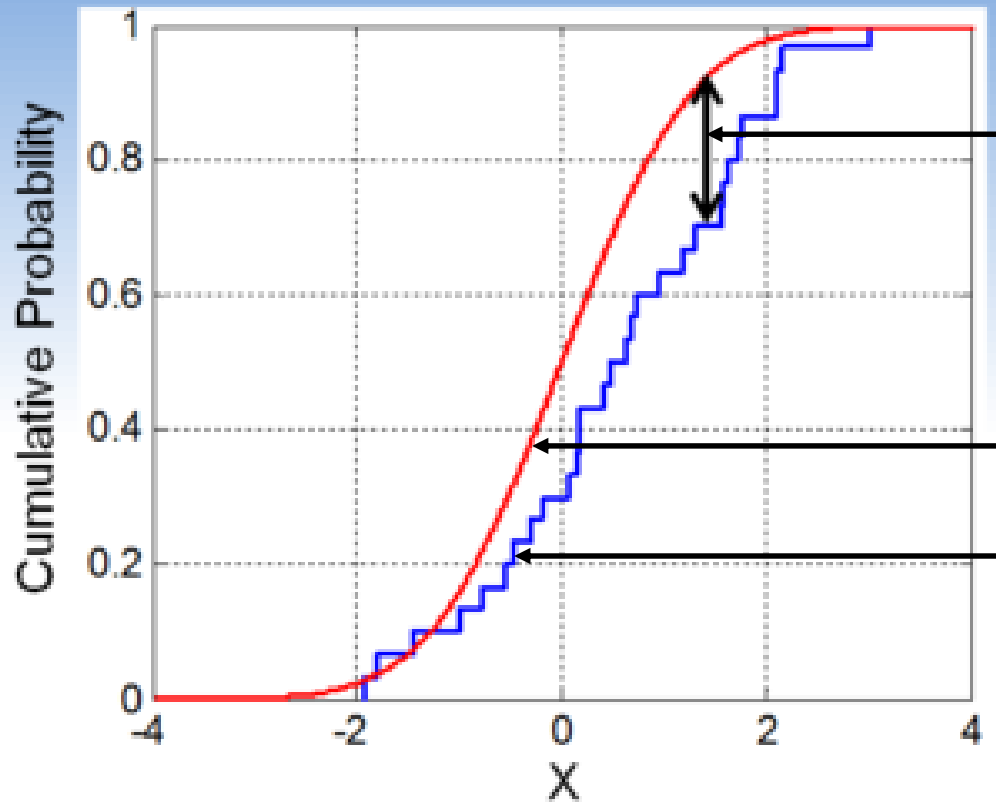
$$\sigma^2(s) = \frac{\phi(s)}{f(s)} = RMSE^2 \frac{g(s)}{f(s)} = RMSE^2 \frac{p(x;\alpha_1,\beta_1)}{p(x;\alpha_2,\beta_2)} \qquad x = s - s_0$$

Let $\beta_1 = \beta_2$, Then:

$$u(s) = \sqrt{\sigma^2(s)} = C * RMSE * (s - s_0)^{(\alpha_1 - \alpha_2)/2}$$

If $\alpha_1 > \alpha_2$, then $u(s)$ is monotonically increasing with s as strongly desired!

Fitting the distribution parameters using cumulative distributions



Kolmogorov-Smirnov (K-S) statistic
Maximum vertical distance between two cumulative distributions

Analytical Cumulative Distribution

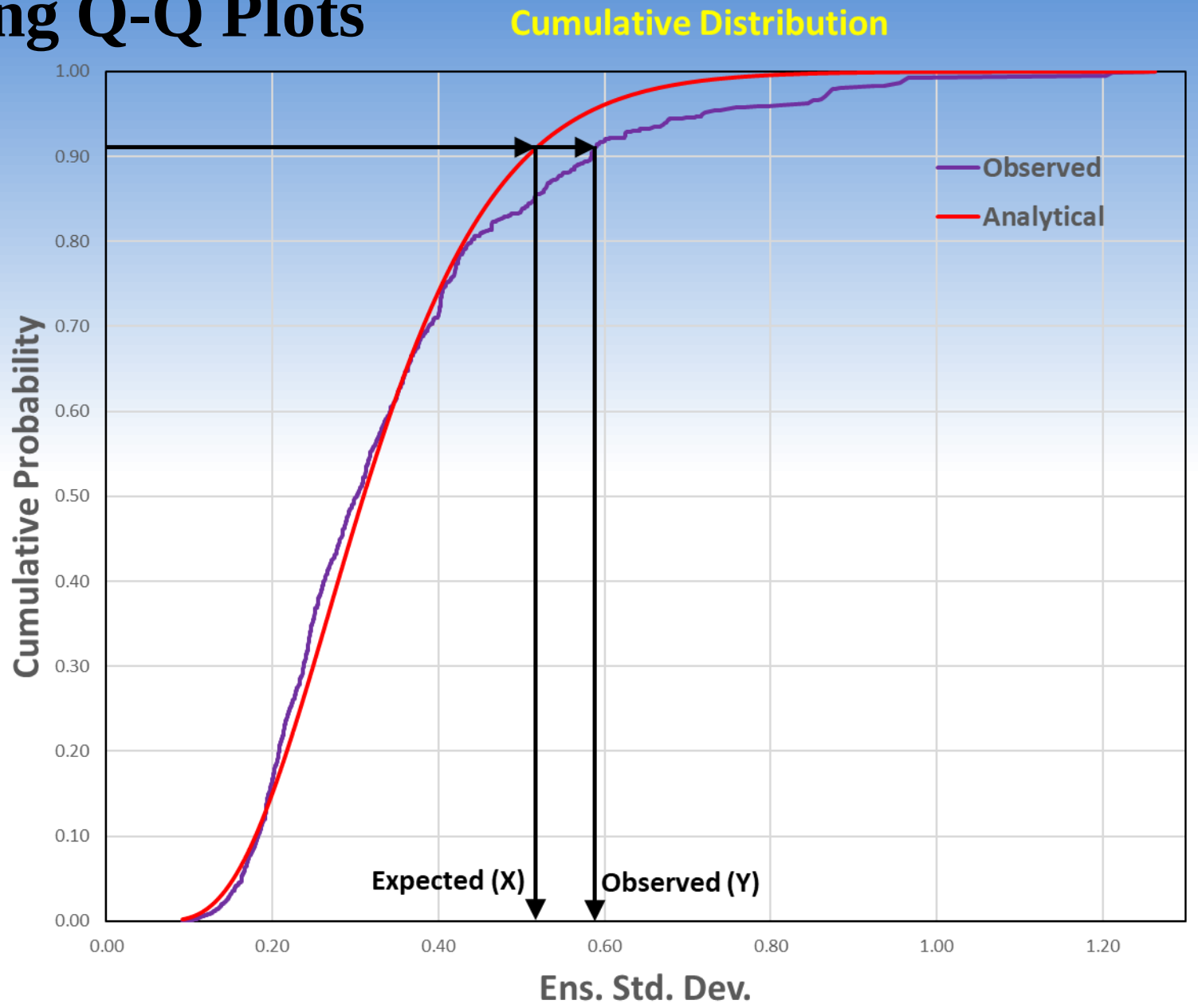
Observed Cumulative Distribution

https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test

Adjust parameters of the analytical distributions to minimize the K-S value

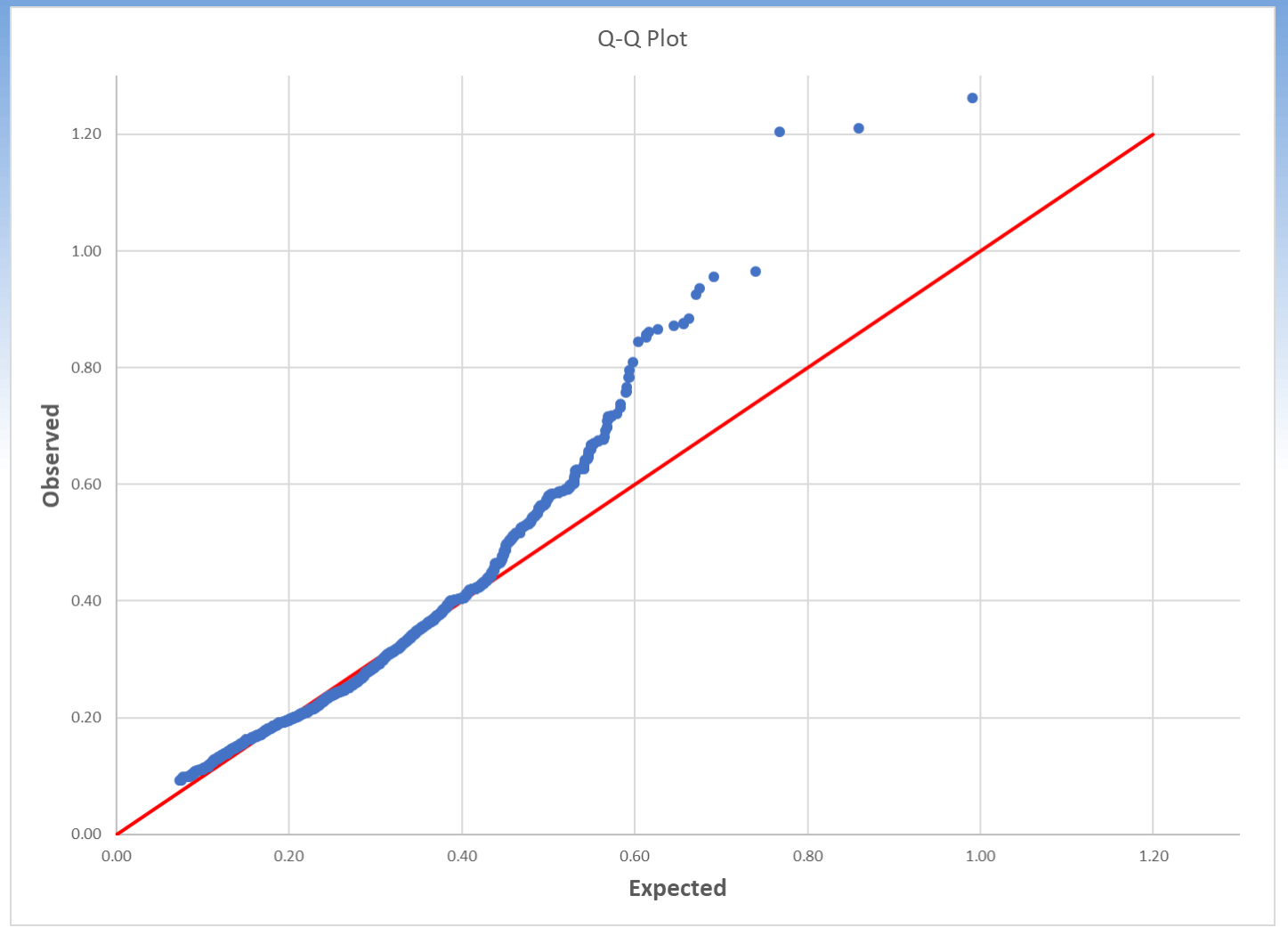
SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

Introducing Q-Q Plots



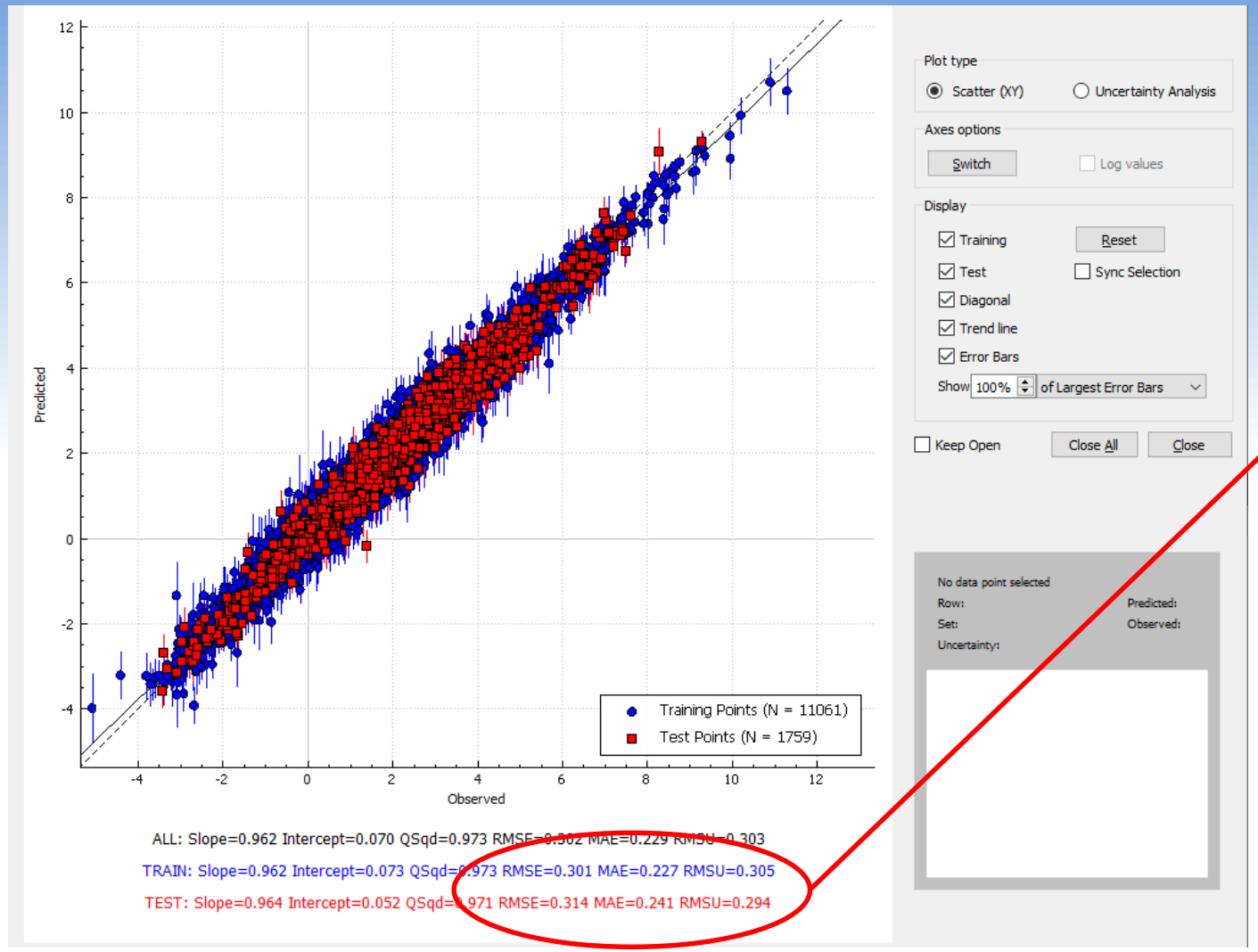
SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

Q-Q Plot



SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

Some Results …



Training Set:
RMSE: 0.301
RMSU: 0.305

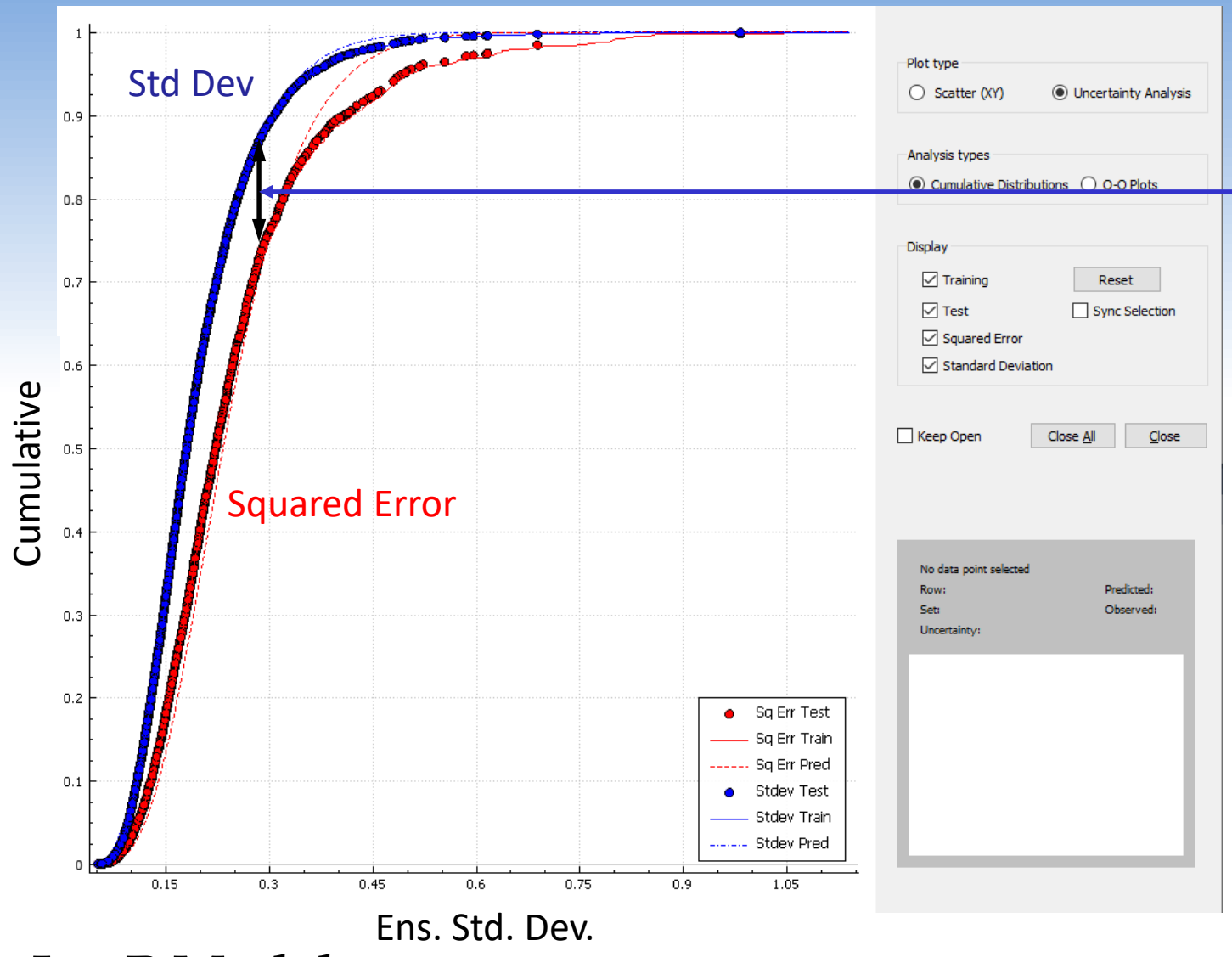
Test Set:
RMSE: 0.314
RMSU: 0.294

RMSU: **Root Mean Squared Uncertainty**

LogP Model

SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

Cumulative Distributions



Vertical separation between cumulatives indicates positive correlation of uncertainty estimate with Ens. Std. Dev.
The greater the separation, the stronger the dependence:

$$u(s) = C * RMSE * (s - s_0)^{(\alpha_1 - \alpha_2)/2}$$

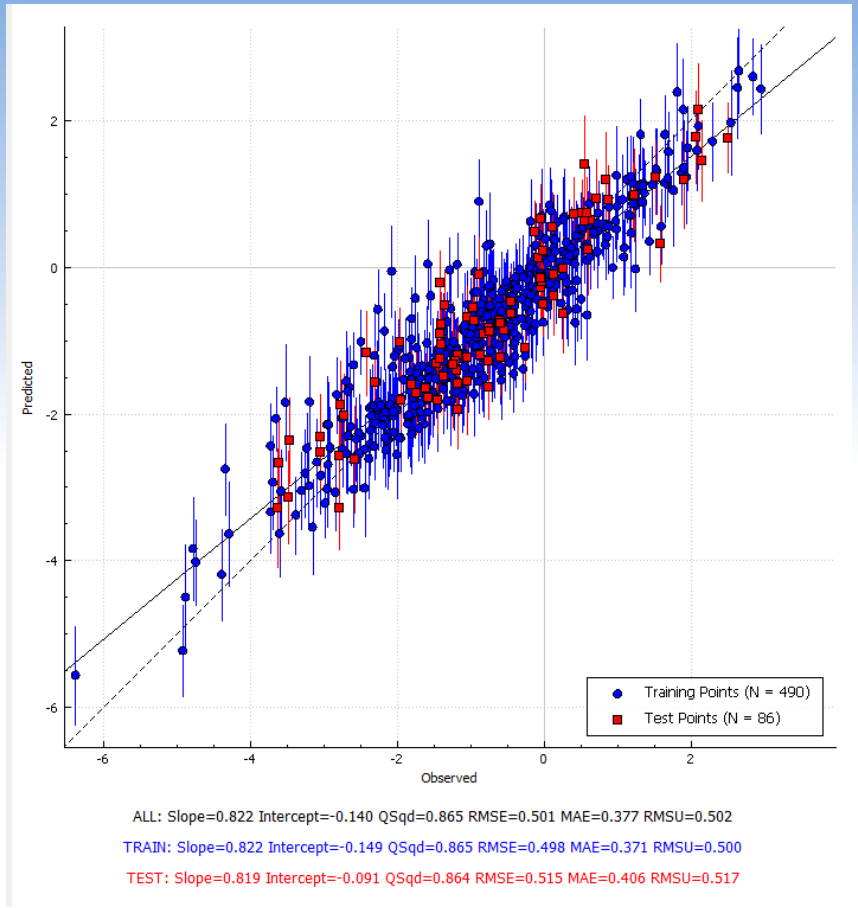
The more Cum Std.Dev. > Cum Sq Err

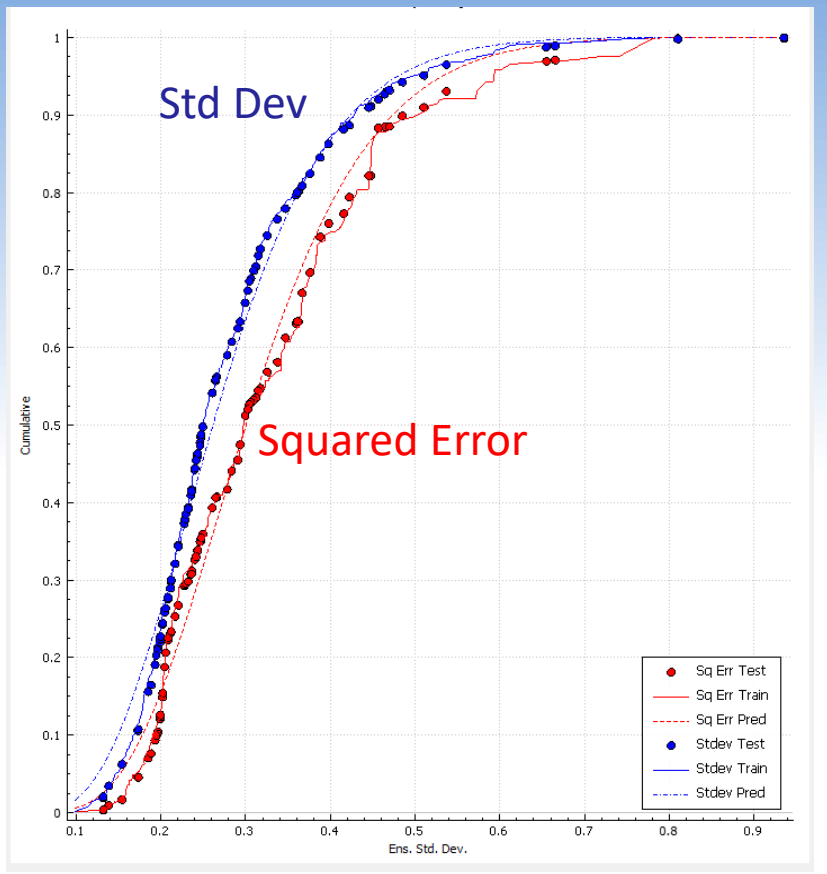
The greater is $\alpha_1 > \alpha_2$

LogP Model

SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

More Results – Fathead Minnow Toxicity





Cumulative Distributions

	Train	Test
RMSE	0.498	0.515
RMSU	0.500	0.517

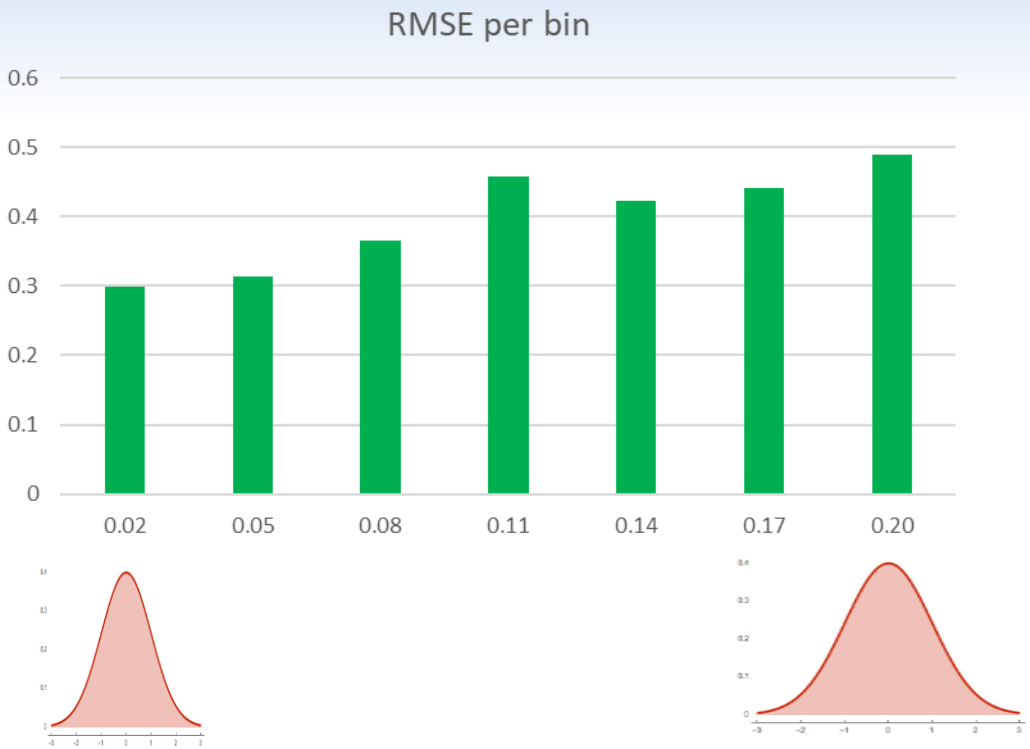
SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

Normalized Error

If each “bin” is a normal distribution with zero mean and std. dev. equal to the uncertainty estimate, then (in the continuous limit) …

Let ε = observed error
Let u = uncertainty
Normalized Error:

$$\rho = \frac{\varepsilon}{u}$$

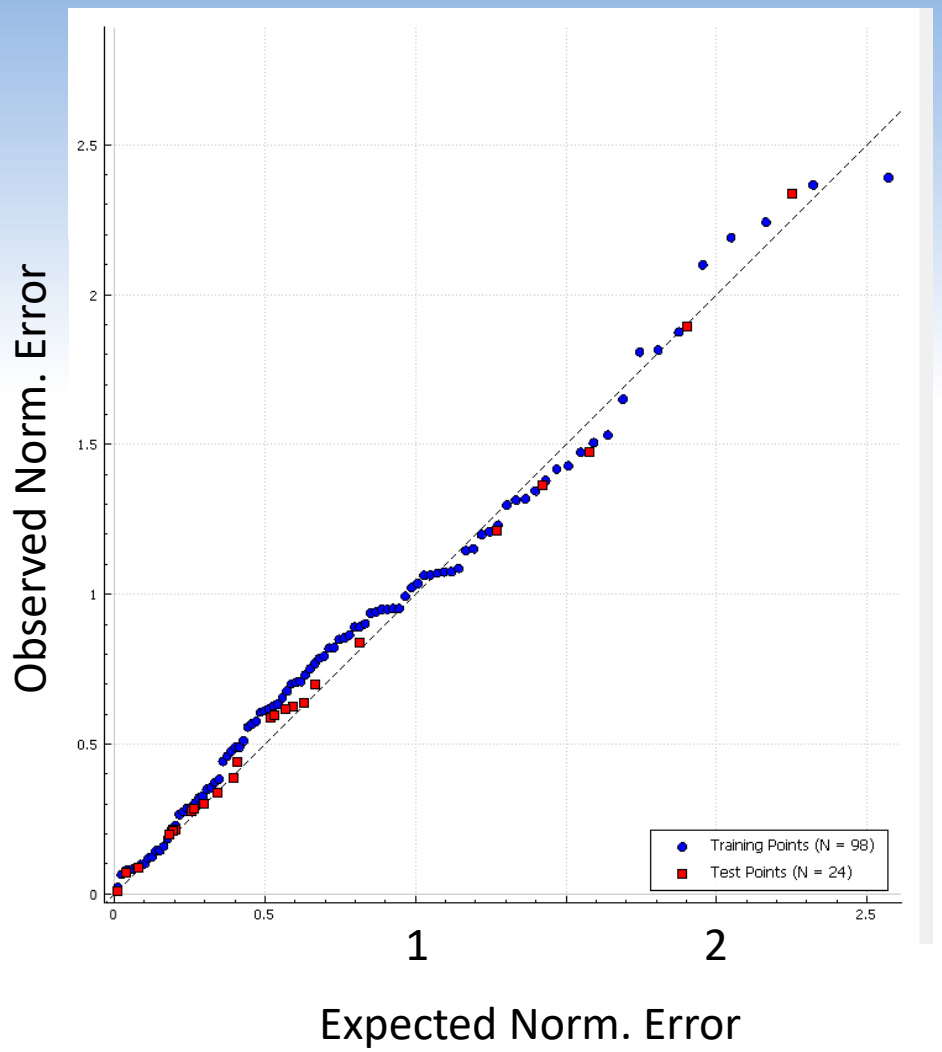


Normalized Error should follow a standard normal distribution with zero mean and unit variance. Its absolute value follows a folded normal distribution. We generate the Q-Q plot for $|\rho|$ compared to the theoretical folded normal distribution. **Note that no parameters are used to fit this Q-Q plot.**

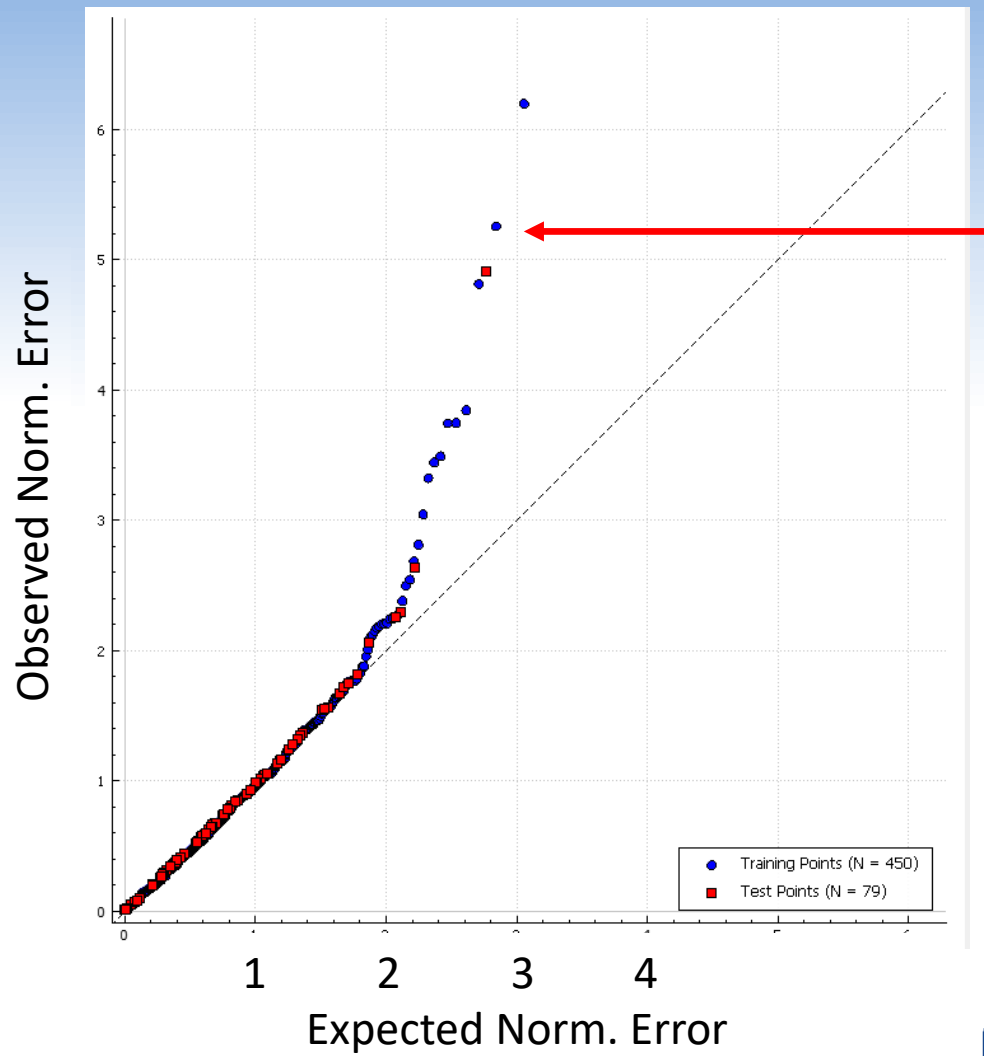
SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

QQ Plots – Normalized Error - Examples

Estrogen Receptor Binding Affinity



Henry's Law Constant – log Space

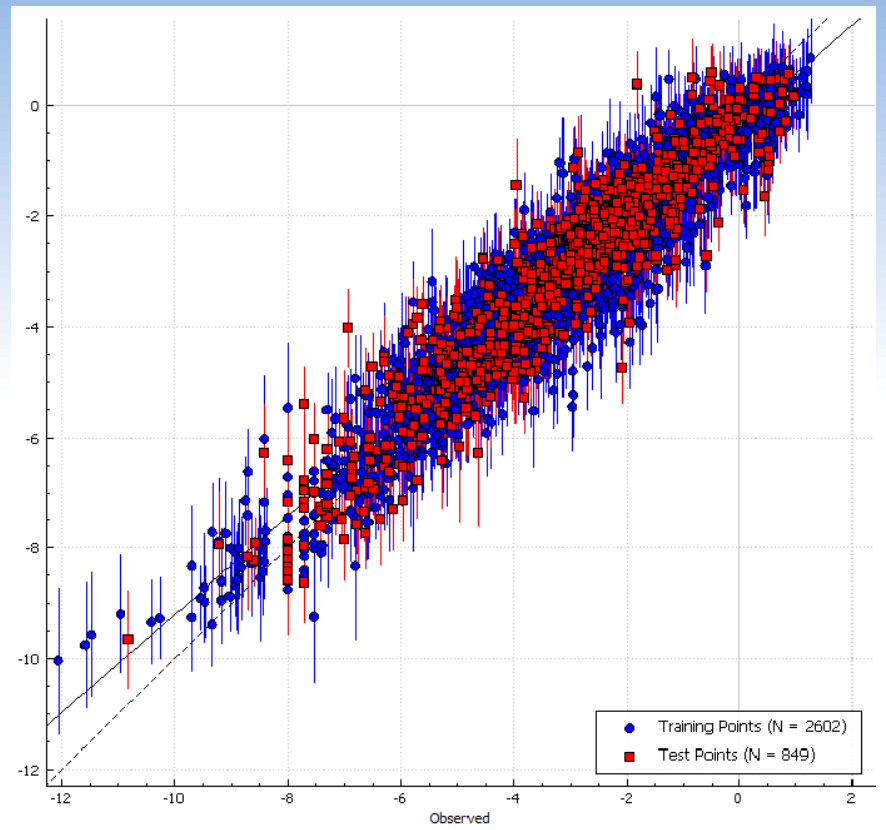


Deviations at $\rho > 2$ indicate a longer tail than standard normal dist. However, ±2 standard deviations represent 95% of a normal dist. So, we expect an uncertainty estimate of $\pm 2\sigma$ to be correct 95% of the time. An uncertainty estimate of $\pm\sigma$ should be correct 68% of the time.

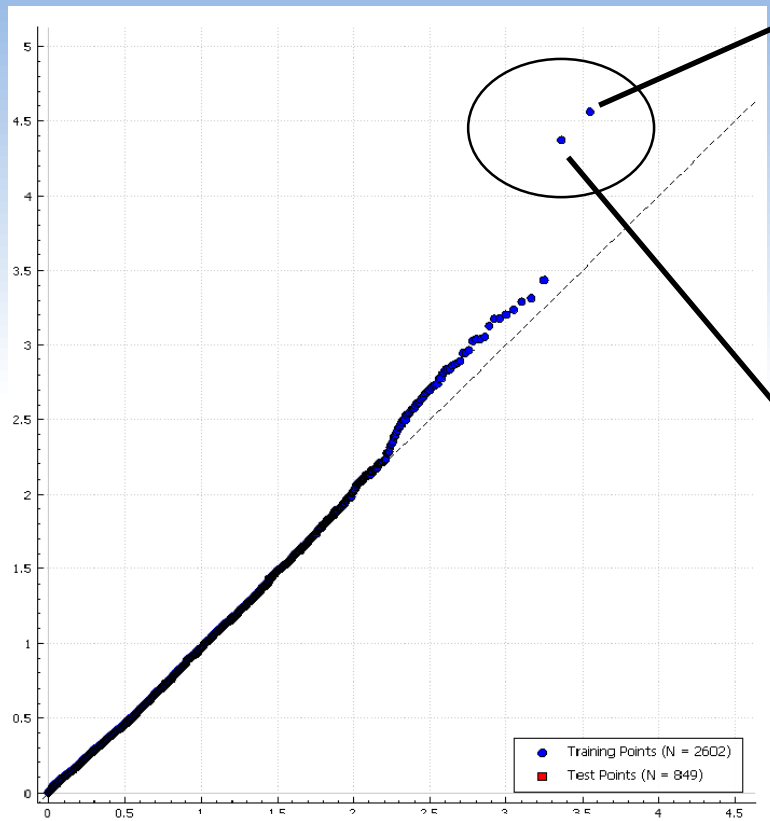
These deviations may also indicate bad data…

SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

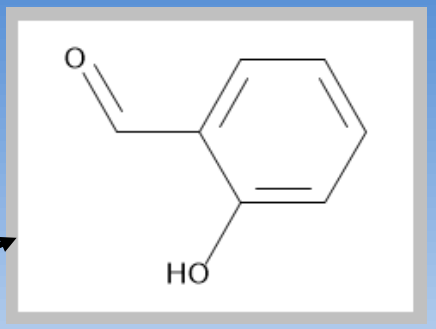
Curating Data with QQ Normalized Error Plots
Solubility Model – log(mol/L)



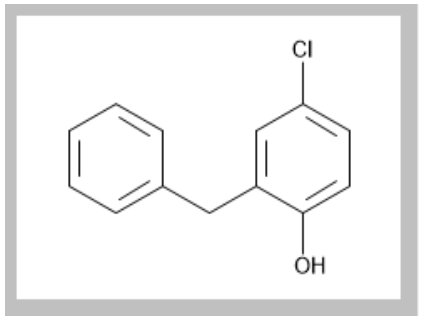
QQ Plot – Normalized Error



1 2



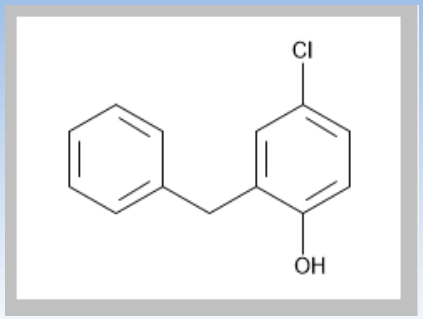
Salicylaldehyde
Predicted: -1.05
Reported: -3.18
Uncertainty: 0.47
Melting Pt.: -7° C (liquid!)



Chlorophene
Predicted: -3.98
Reported: -1.72
Uncertainty: 0.52

SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

Digging Deeper …



Chlorophene
Reported (Aquasol): -1.72 (log units) → 1.9 E-2 mol/L
“Estimated from graph”
Allawala and Riegelman, J Am. Pharm. Assoc. XLII, 267 (1953)
More recently:
EPA: 6.81 E-4 mol/L
pubchem.ncbi.nlm.nih.gov/compound/2-Benzyl-4-chlorophenol

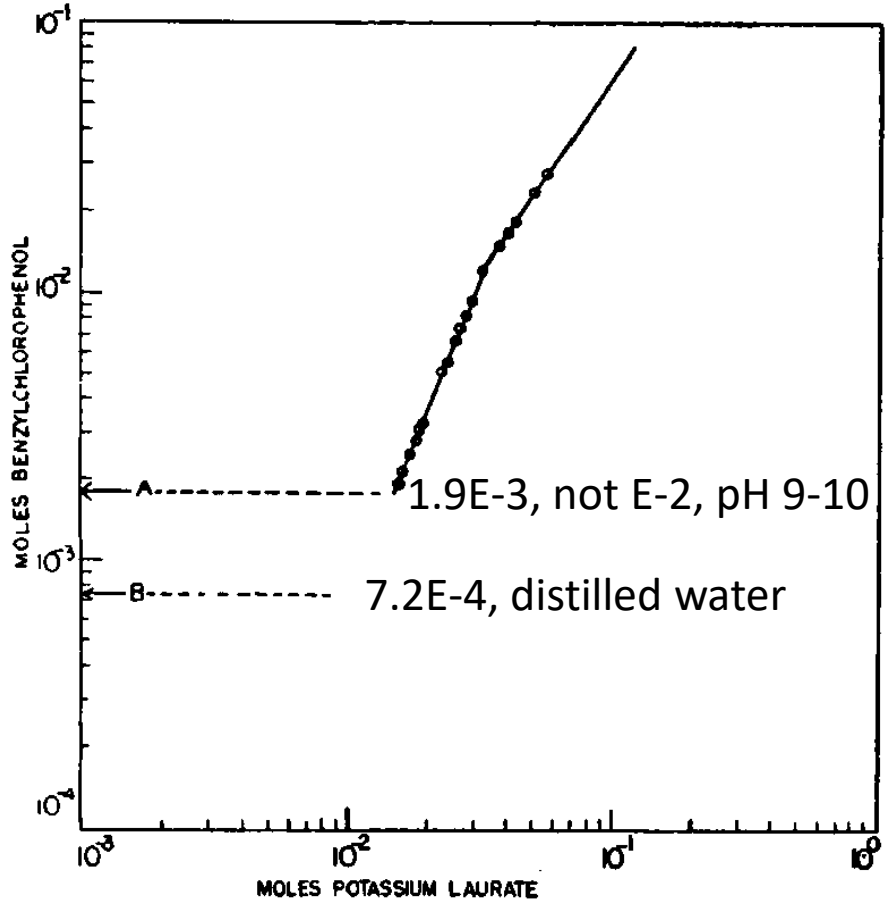
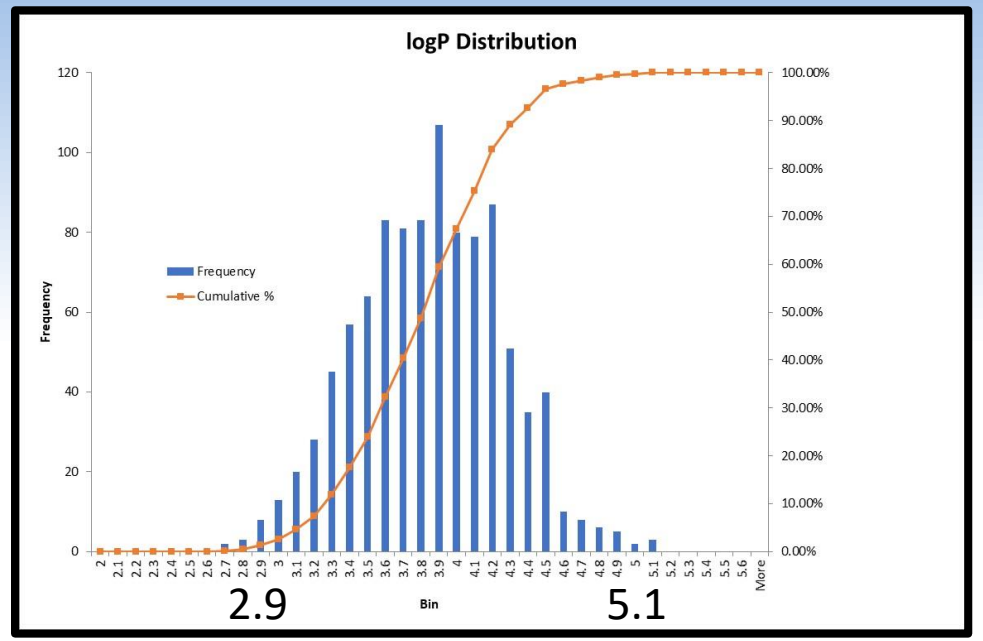


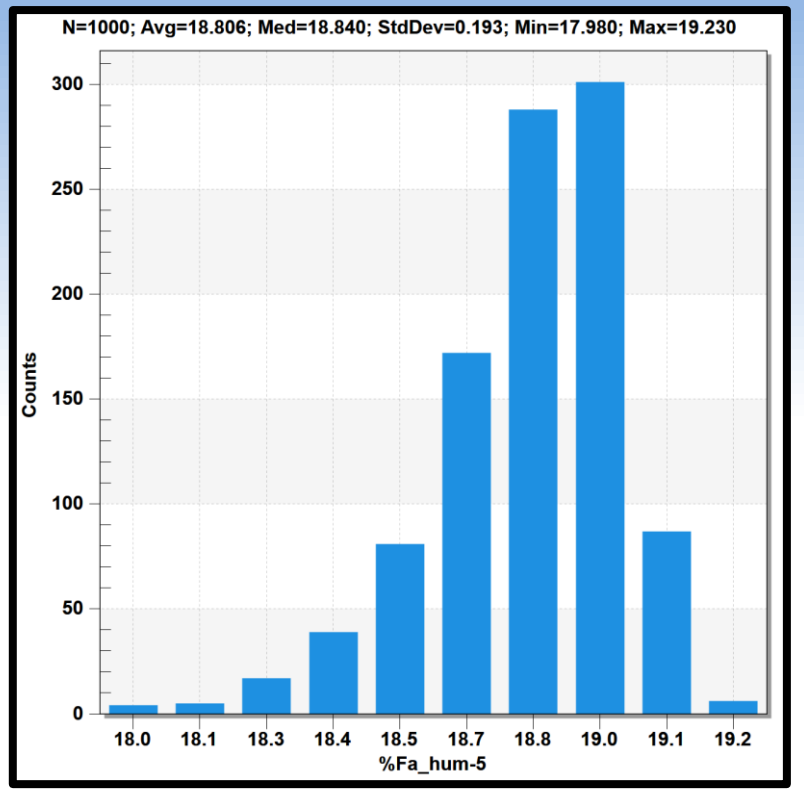
Fig. 8.—A log–log plot of the solubilization of benzylchlorophenol (5-chloro-2-hydroxy diphenylmethane) in moles/L. by solution in potassium laurate (moles/L.) at 20°. A: saturation solubility of benzylchlorophenol at *p*H 9–10; B: saturation solubility of benzylchlorophenol in distilled water.

SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

Using uncertainty estimates in pharmacokinetic simulations



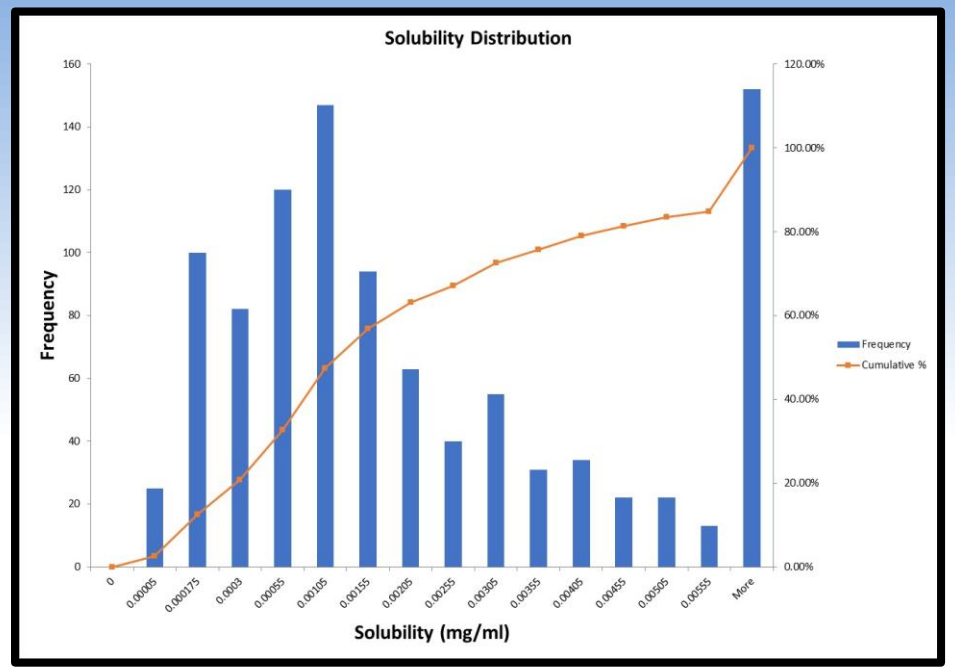
Create a randomly sampled normal distribution of 1000 logP values using uncertainty estimate as std. dev.



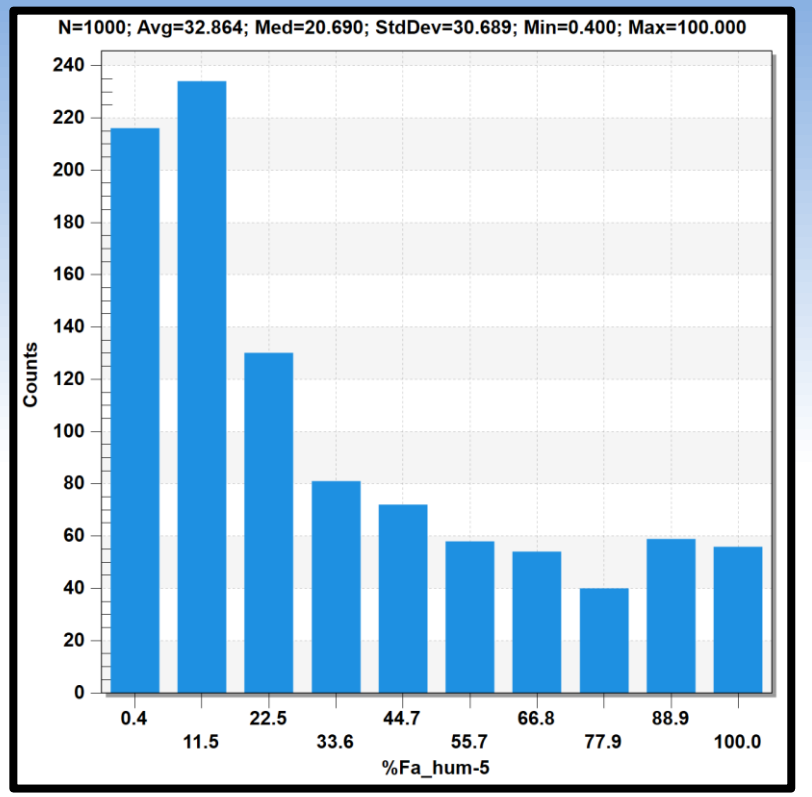
%Fraction absorbed shows little sensitivity to logP range.

SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

Using uncertainty estimates in pharmacokinetic simulations - part 2



5E-5 5.E-3



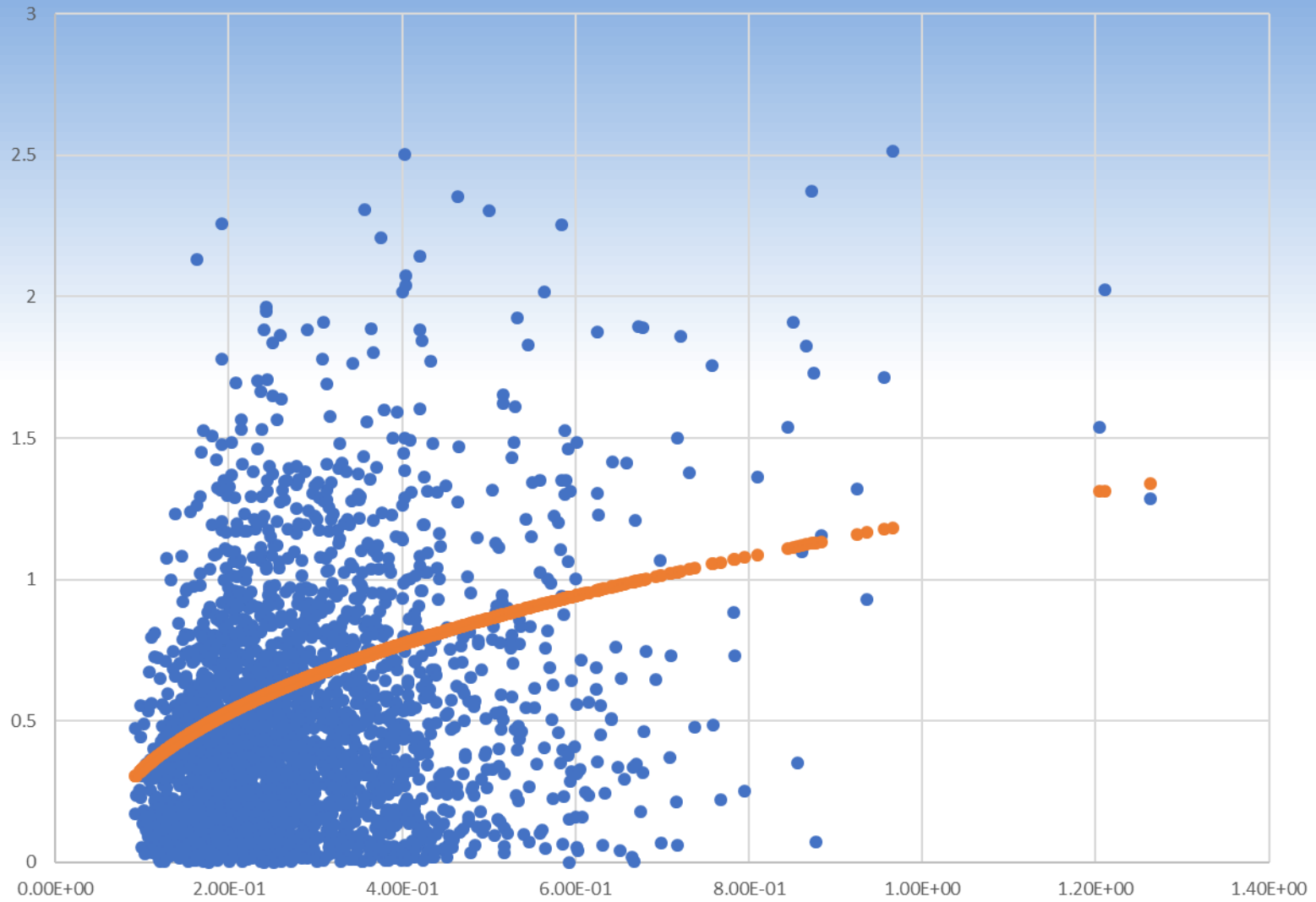
0.4 100

Create a randomly sampled log-normal distribution of 1000 Sw values using uncertainty estimate as std. dev.

%Fraction absorbed shows high sensitivity to Sw range. Solubility prediction has low confidence – should probably be measured!

SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

At final glance …



Is there signal in this noise?
Yes!

For this data set,
uncertainty is approx.
square root function,
not linear or exponential

SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

Conclusions

- Left as an exercise for the reader …

Acknowledgements

- Robert D. Clark (co-author)
- Pankaj Daga
- Michael Lawless
- David Miller