

Estimating Predictive Uncertainty for Ensemble Regression Models by Gamma Error Analysis

Bob Clark & Marvin Waldman

Simulations Plus, Inc.

Lancaster CA USA

bob@simulations-plus.com

EuroQSAR 2018

The Standard Error (SE) of Prediction: a Measure of Individual Uncertainties



PRO:

- Is easy to display & to grasp visually
- Supports easy comparisons between predictions
- Allows confidence intervals to be calculated at an arbitrary level of uncertainty
- Everybody knows what it means

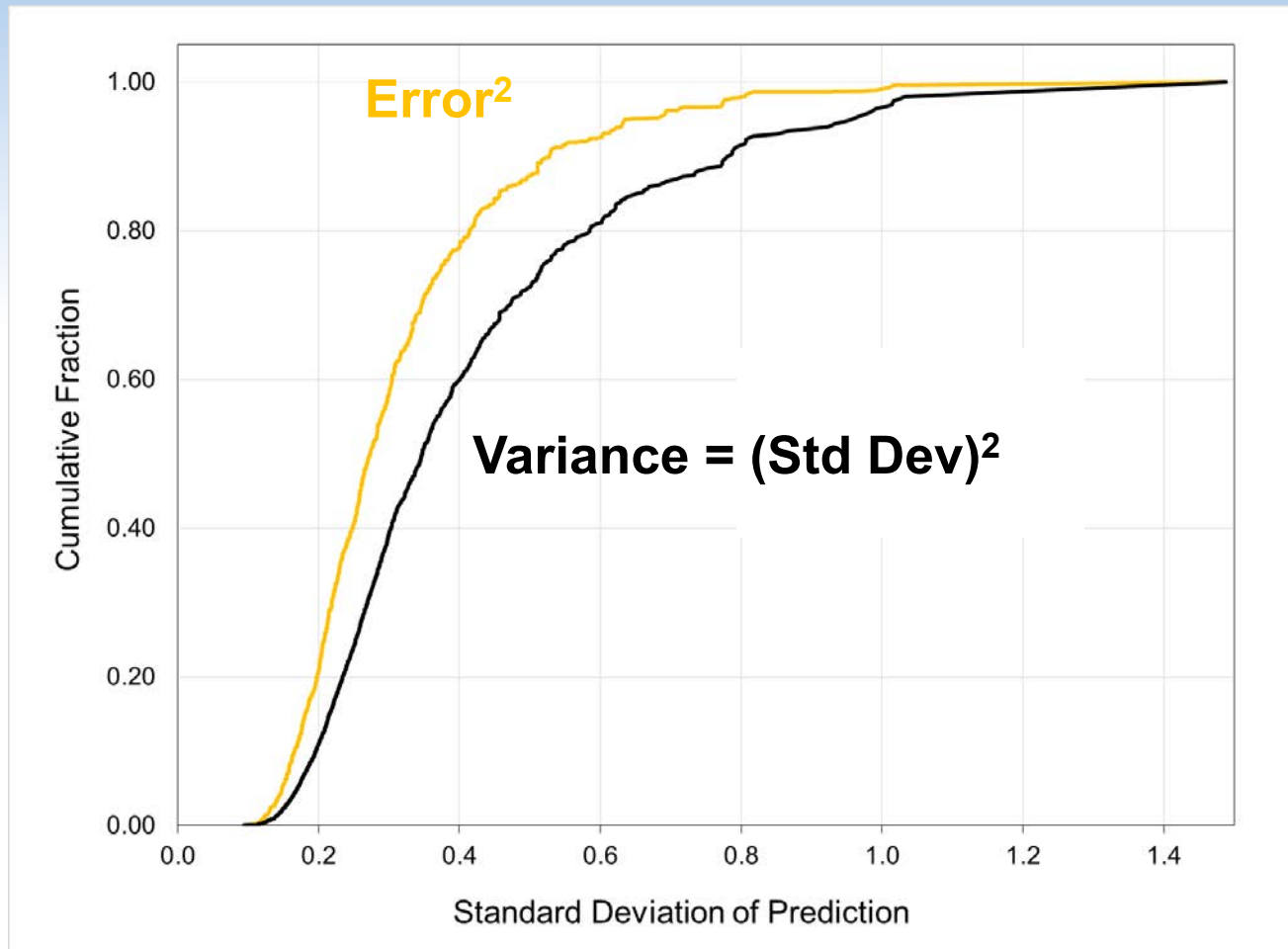
CON:

- To be *precise*, predictive errors must follow an approximately normal distribution with a standard deviation $\sigma_i = SE_i$
- To be *useful*, predictive error should follow an approximately normal distribution with a standard deviation $\sigma_i \leq SE_i$
- It cannot account for non-random error (bias) in predictions
- Nobody knows what it means

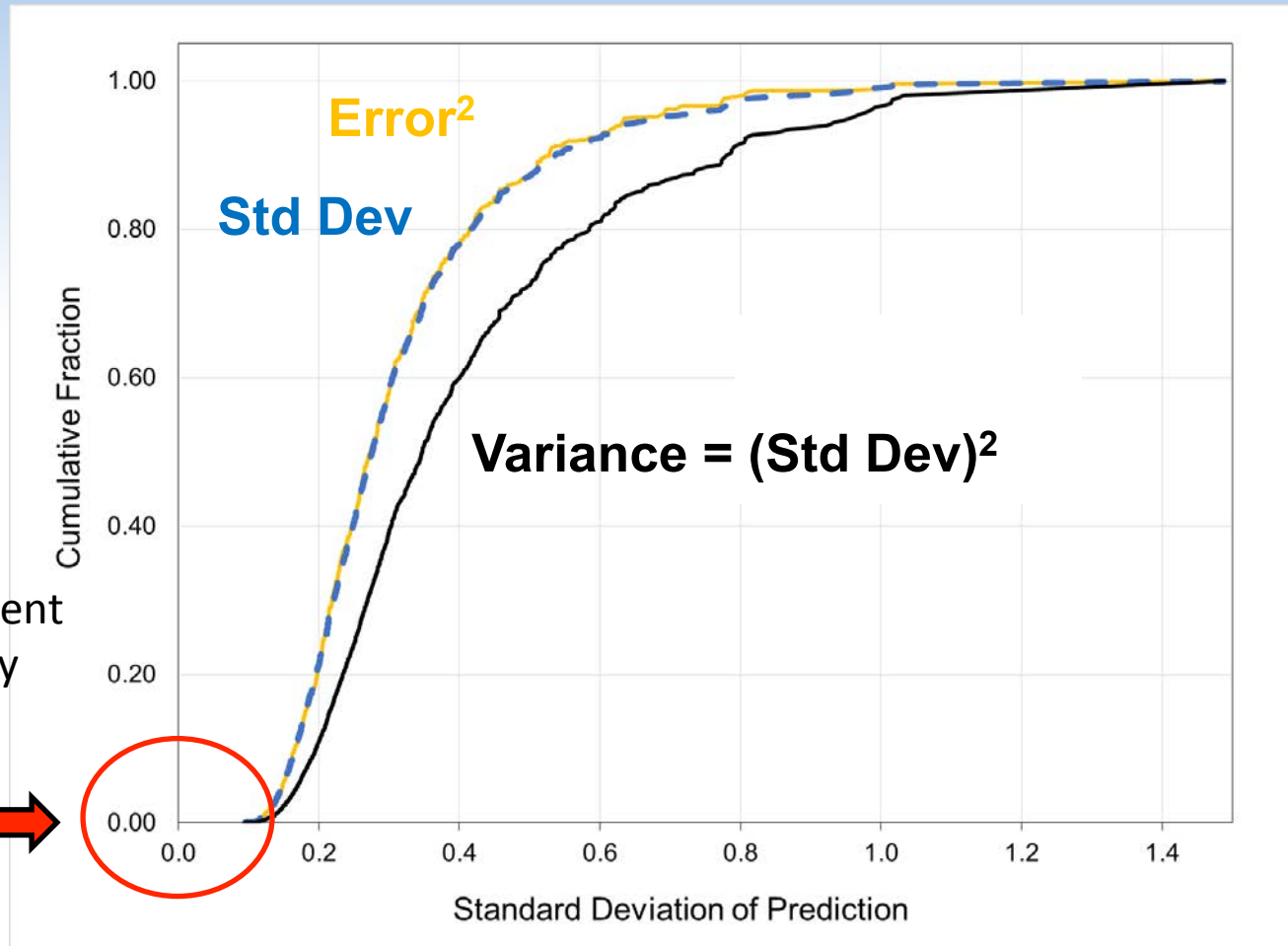
Some Relevant Previous Work on Ensemble Predictivity

- 
- B. Beck *et al.* *J Chem Inf Comput Sci* **2000**, 40, 1046-1051
 - used the variance in artificial neural net ensembles to estimate uncertainty
 - S. Weaver & M.P. Gleeson. *J Molec Graph Model* **2008**, 26, 1315–1326
 - estimated accuracies of individual regression predictions
 - U. Sahlin *et al.* *Mol Inf* **2011**, 30, 551 – 564
 - uncertainty and risk assessment
 - S. Modi *et al.* *J Comput-Aided Mol Des* **2012**, 26, 1017-1033
 - consensus models for *in silico* Ames testing
 - R.P. Sheridan. *J Chem Inf Model* **2012**, 52, 814–823
 - using variance across random forest predictions to help assess confidence
 - C.E. Keefer *et al.*, *J Chem Inf Model* **2013**, 53, 368–383
 - confidence metric based on nearest neighbor consensus
- 
- R.D. Clark *et al.* *J Cheminfo* **2014**, 6, 34.
 - Using beta binomials to estimate classification uncertainty.

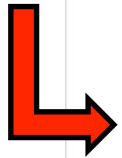
The “Aha!” Moment that Led Us to Gamma Error Analysis



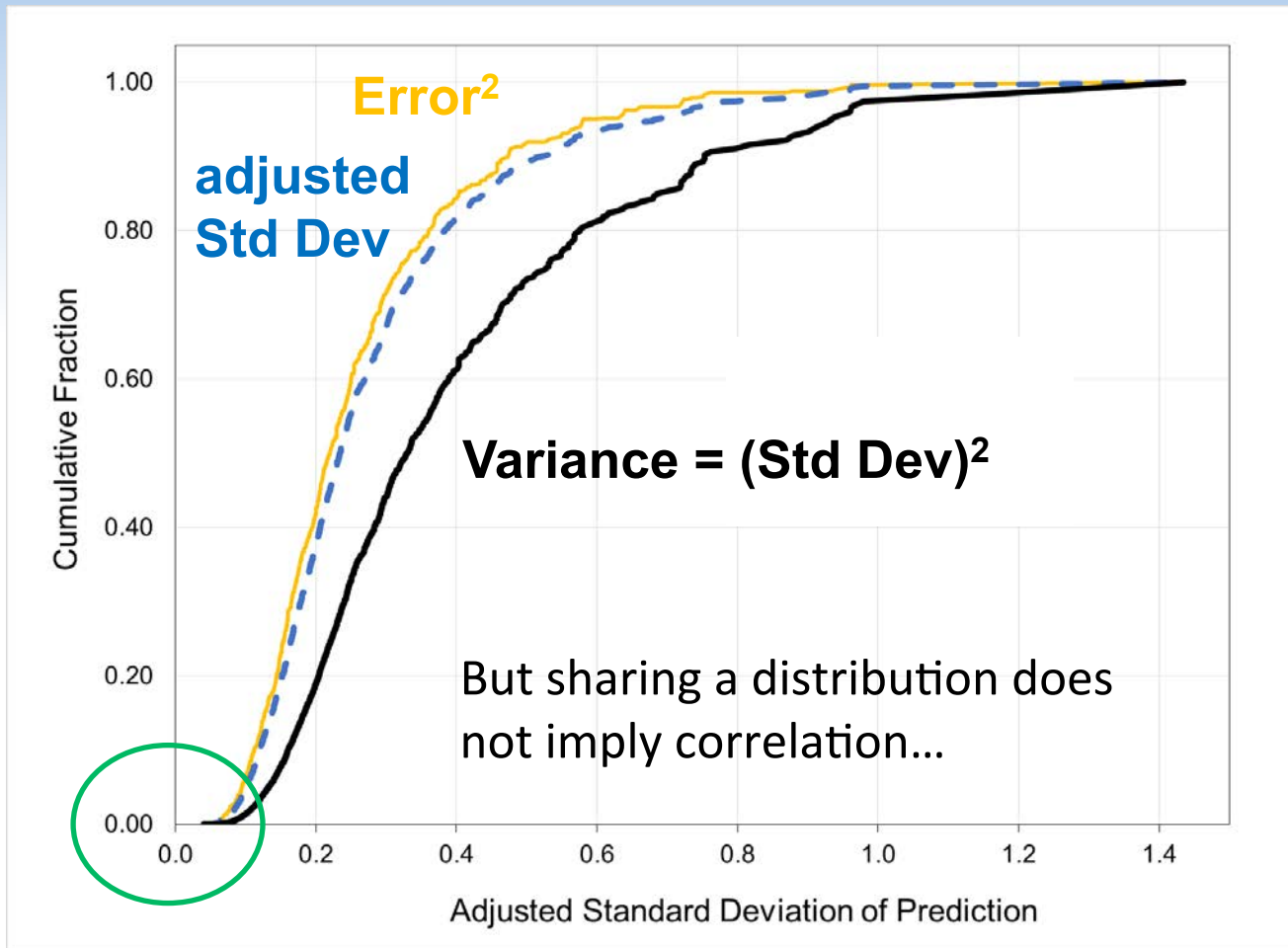
The “Aha!” Moment that Led Us to Gamma Error Analysis



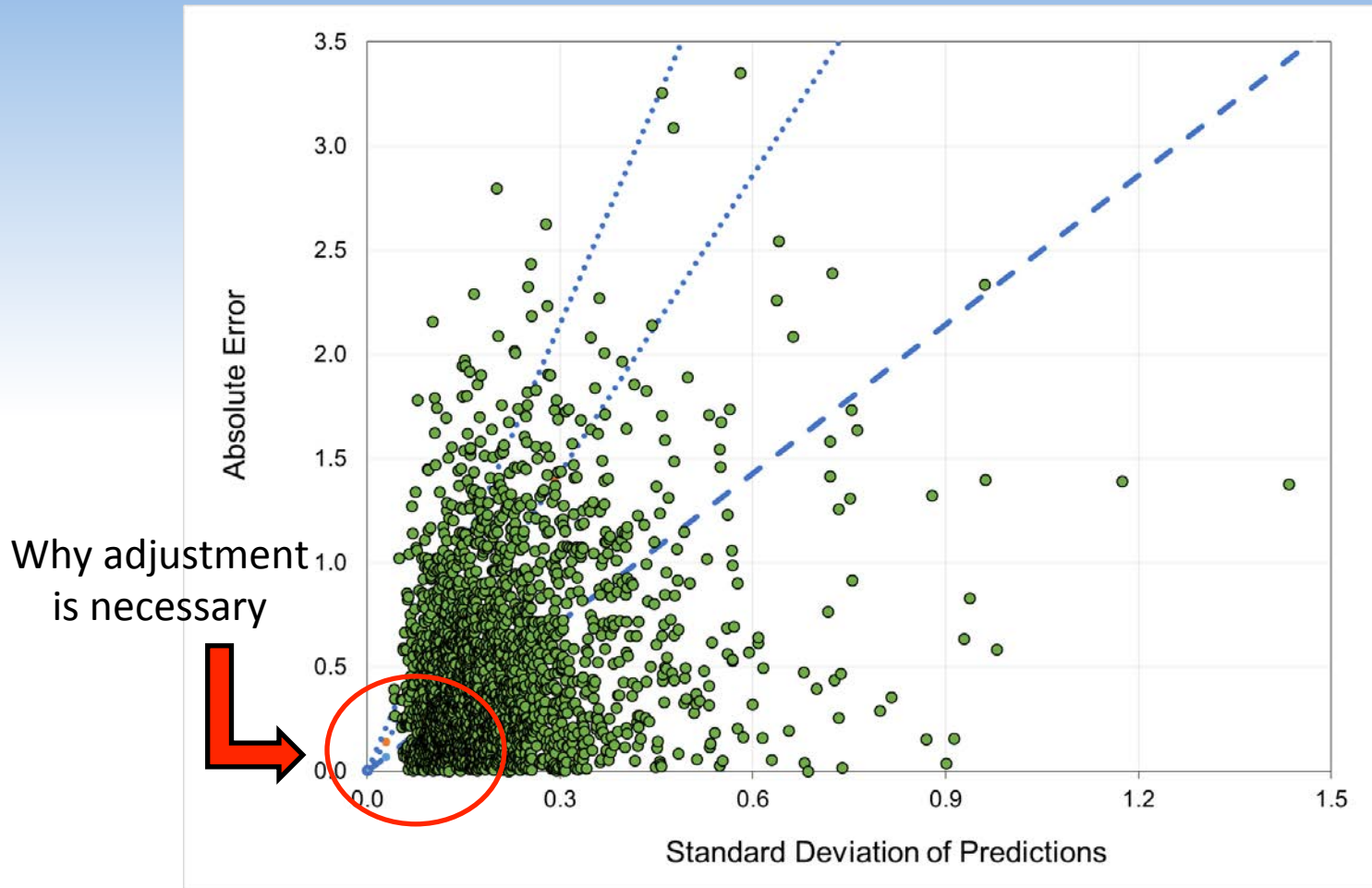
Why adjustment
is necessary



The “Aha!” Moment that Led Us to Gamma Error Analysis



The Relationship Is Not Obvious...

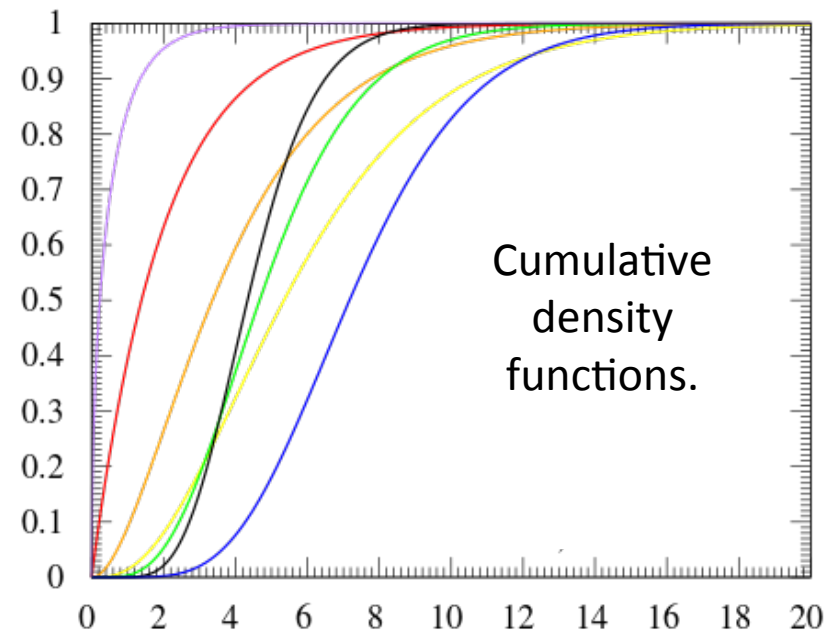
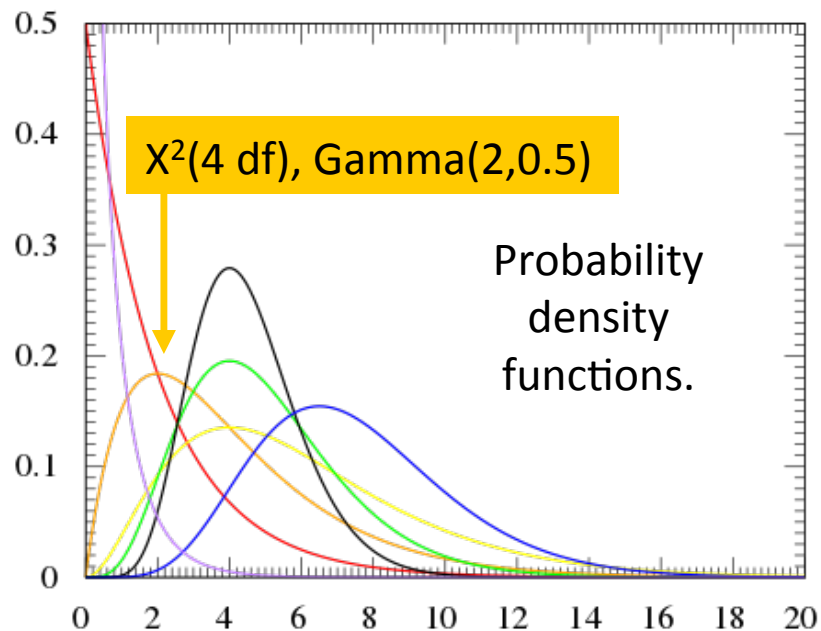


...and is not necessarily real. It is a somewhat subjective matter of perspective.

Generality of Gamma Error Analysis

- The specific examples presented are for artificial neural network ensemble (ANNE) regression models as implemented in the ADMET Modeler™ Module of ADMET Predictor.™
 - each ensemble consists of 33 networks that are trained separately
 - each network has a single hidden layer and single output
 - All networks in an ensemble have the same number of inputs and the same number of neurons
 - each network is trained uses a random verification subset of the training pool to protect against early stopping
 - An external test set is held out before supervised training begins
- That said, the method is based on fundamental statistical principles. Hence there is reason to expect it to be broadly applicable to ensemble models where the constituent submodels are trained against a common task.

Meet the Gamma Distribution

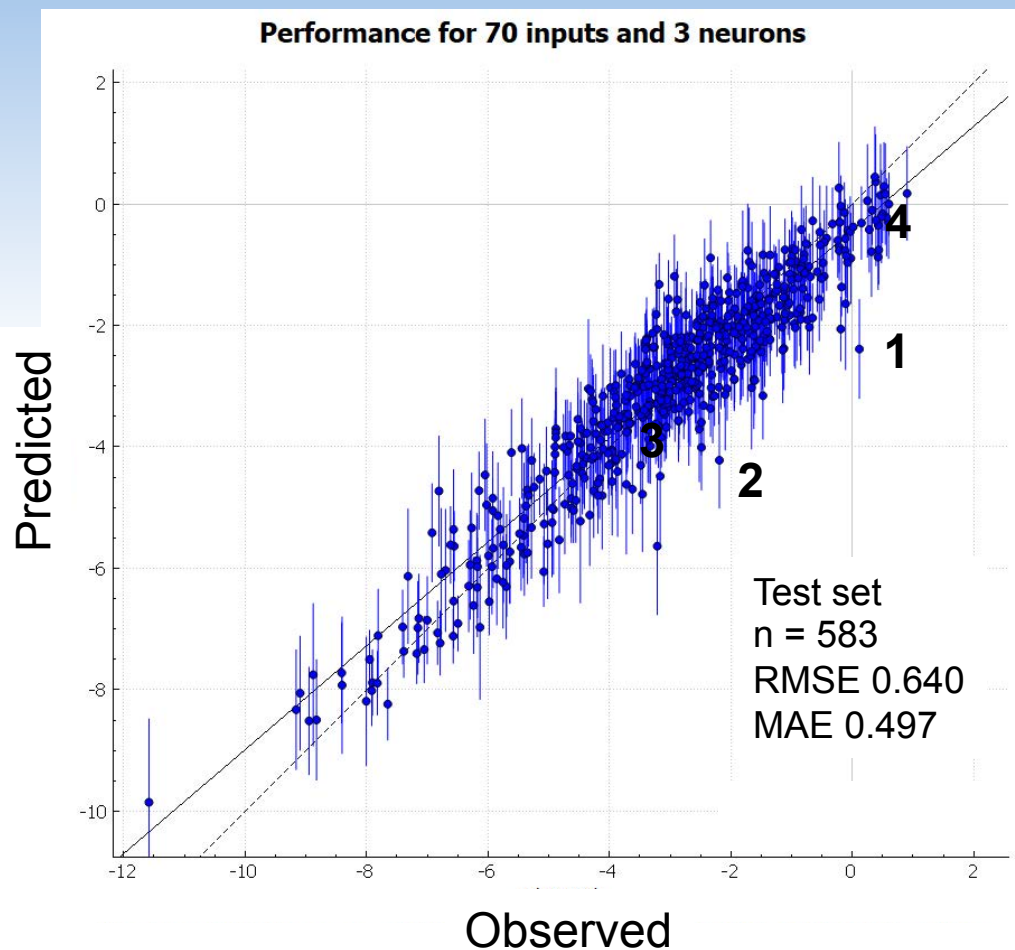


The variance (squared standard deviation) of a random sample from a standard normal distribution follows a chi square (χ^2) distribution, which is a special case of the gamma distribution where α is a half-integer and $\beta = 0.5$.

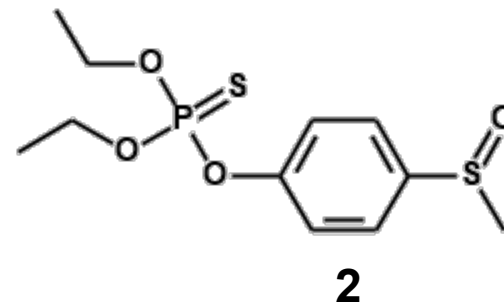
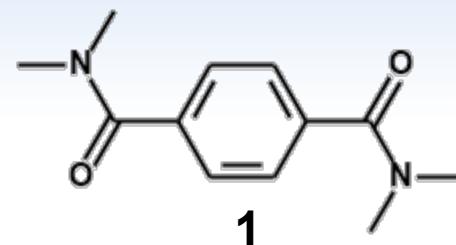
How Is a Gamma Error Analysis Done?

1. Calculate the mean and standard deviation SD_i of the predictions generated by the individual submodels for each observation i in the **training pool**. The average becomes the ensemble prediction $pred_i$. The error err_i is the difference between the predicted and the observed value.
2. Simultaneously fit two cumulative gamma distributions to the squared errors and SDs for the training pool predictions:
 - $G(SD_{adj}; \alpha_{SD}, \beta)$ and $F(err^2, SD_{adj}; \alpha_{err}, \beta)$, with $SD_{adj} = SD_i - SD_0$
3. Calculate the estimated standard error for $pred_q$ (SE_q) for a new instance q from the gamma density functions g and f and the overall root mean square error (RMSE):
 - $SE_q^2 = RMSE^2 \times f(SD_q - SD_0) / g(SD_q - SD_0)$
4. Compare observed errors in the **test set** to their estimated SEs.

Example: ANNE Model* for Log S (M)

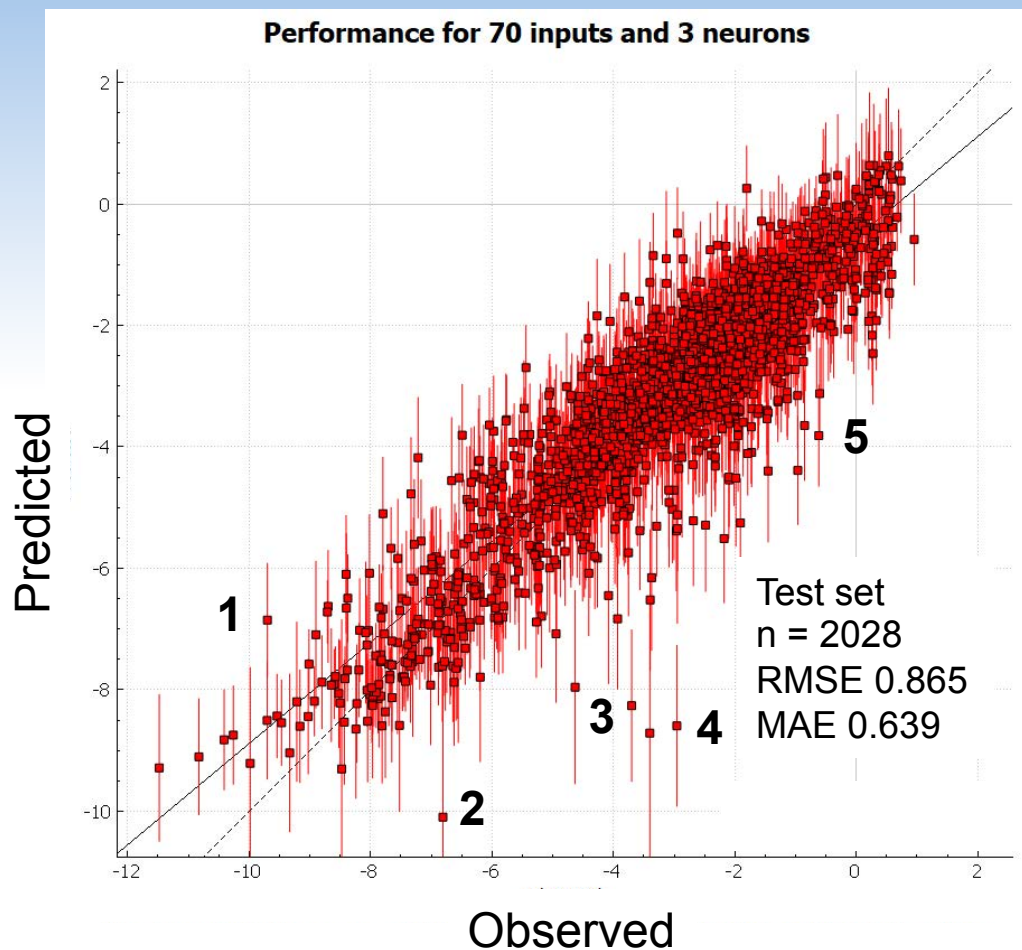


Underpredicted solubilities:

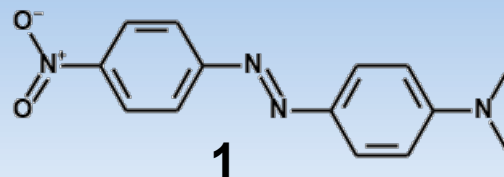


*Built in ADMET Modeler™ in ADMET Predictor™ 9.x (dev)

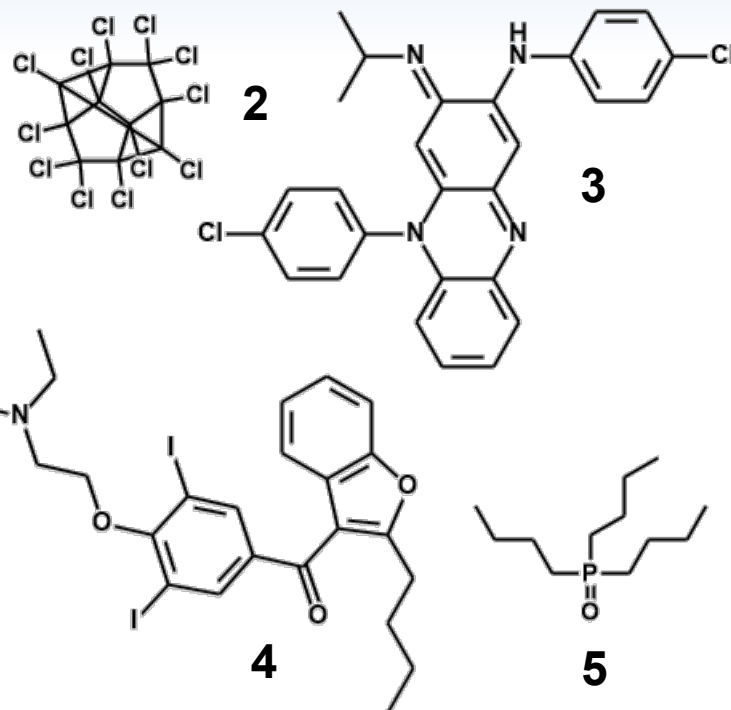
ANNE Model for Log S (M): LARGE Test Set



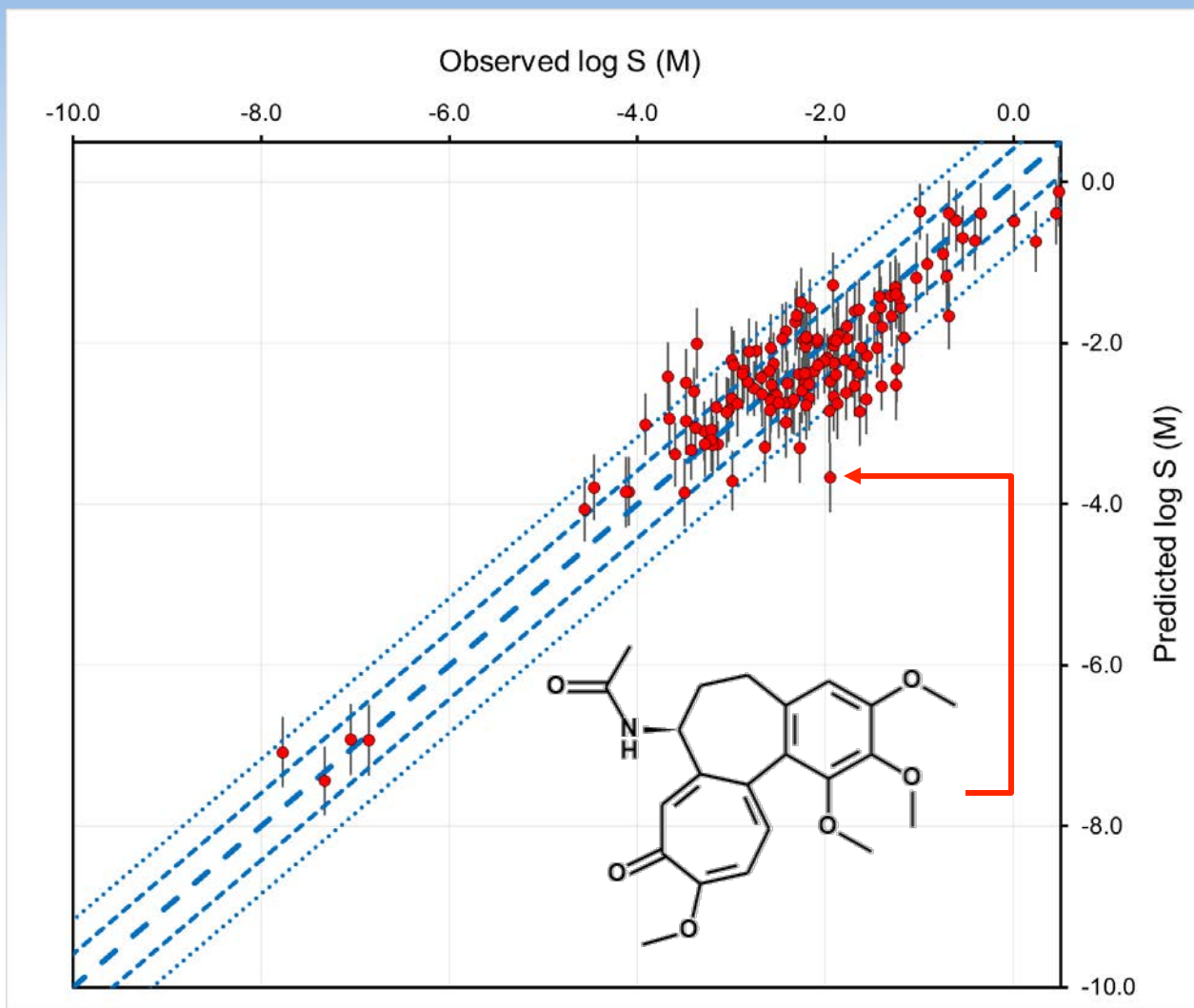
Overpredicted solubility:



Underpredicted solubilities:



Lowest 200 SE Estimates for the Test Set

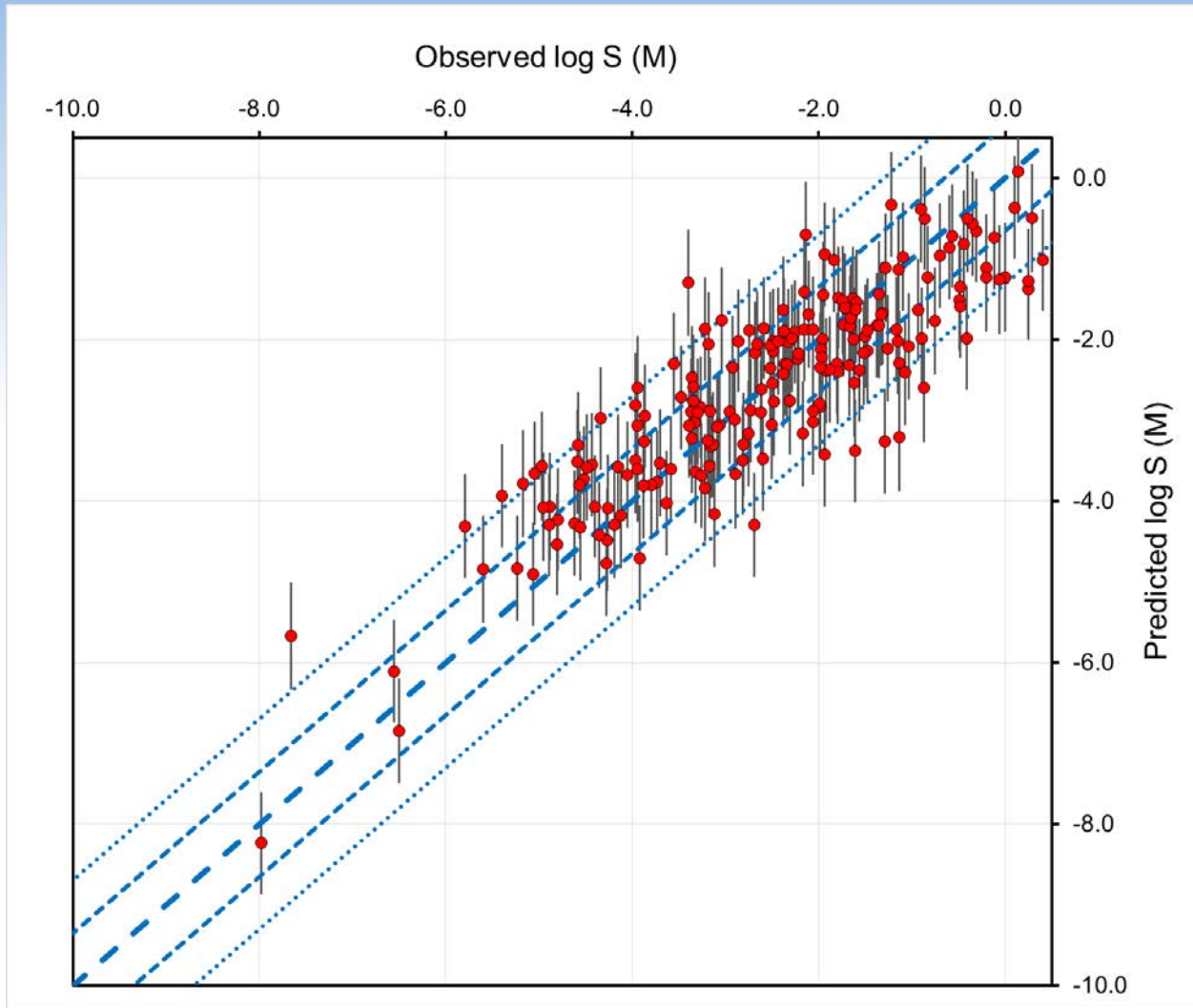


Best 200 Test
RMS SE_i (est) 0.420
RMSE (obs) 0.478

Expected distribution:
68 : 27 : 5%

Obs. Distribution:
54.5 : 28.5 : 17%

200 Middle-ranking SEs from the Test Set

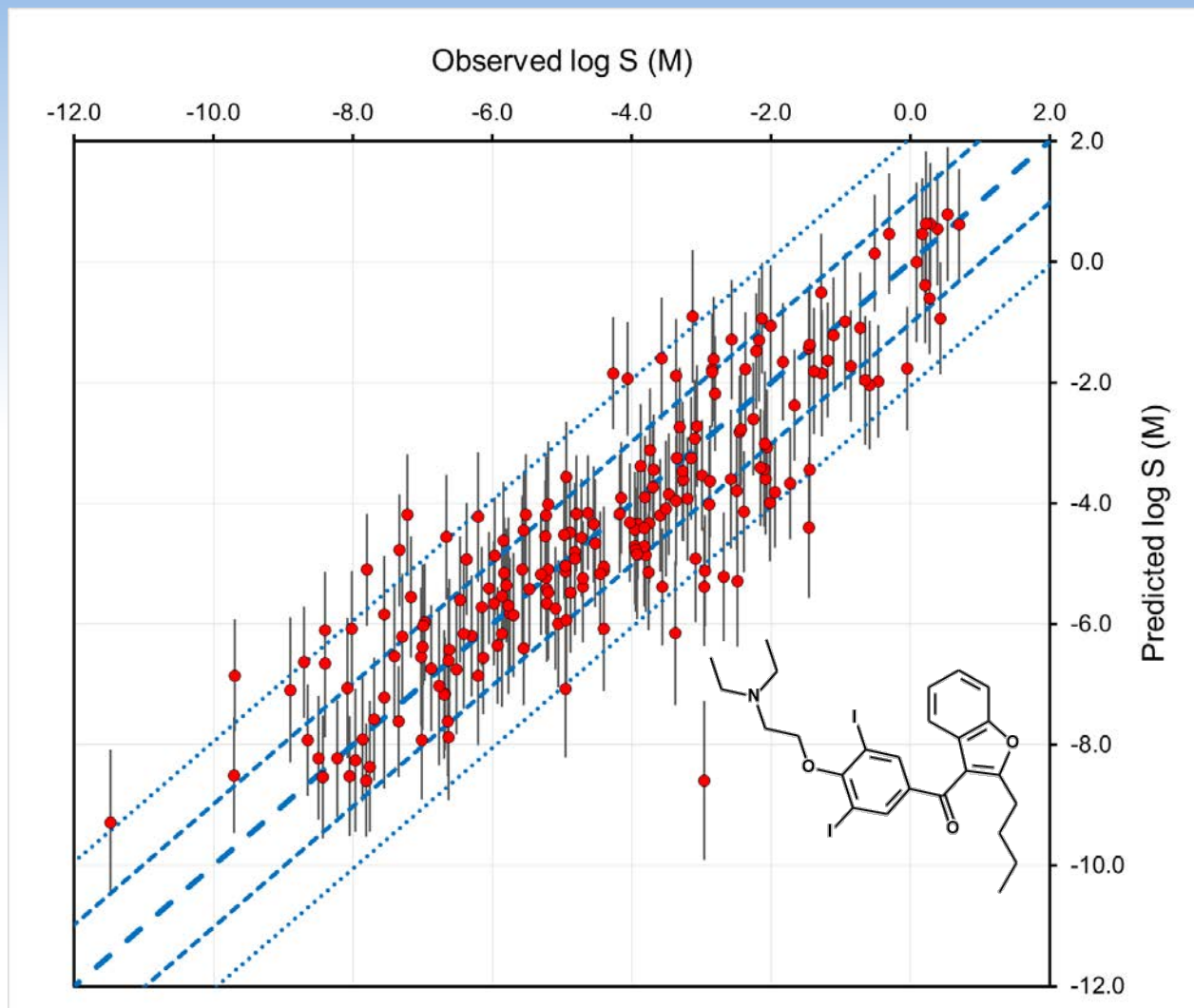


Middle 200 Test
RMS SE_i (est) 0.651
RMSE (obs) 0.770

Expected distribution:
68 : 27 : 5%

Obs. Distribution:
60.5 : 28 : 11.5%

200 Highest Test Set SE Estimates

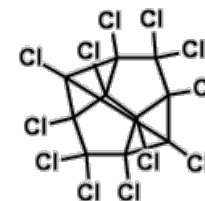


Worst 200 Test
 RMS SE_i (est) 1.025
 RMSE (obs) 1.186

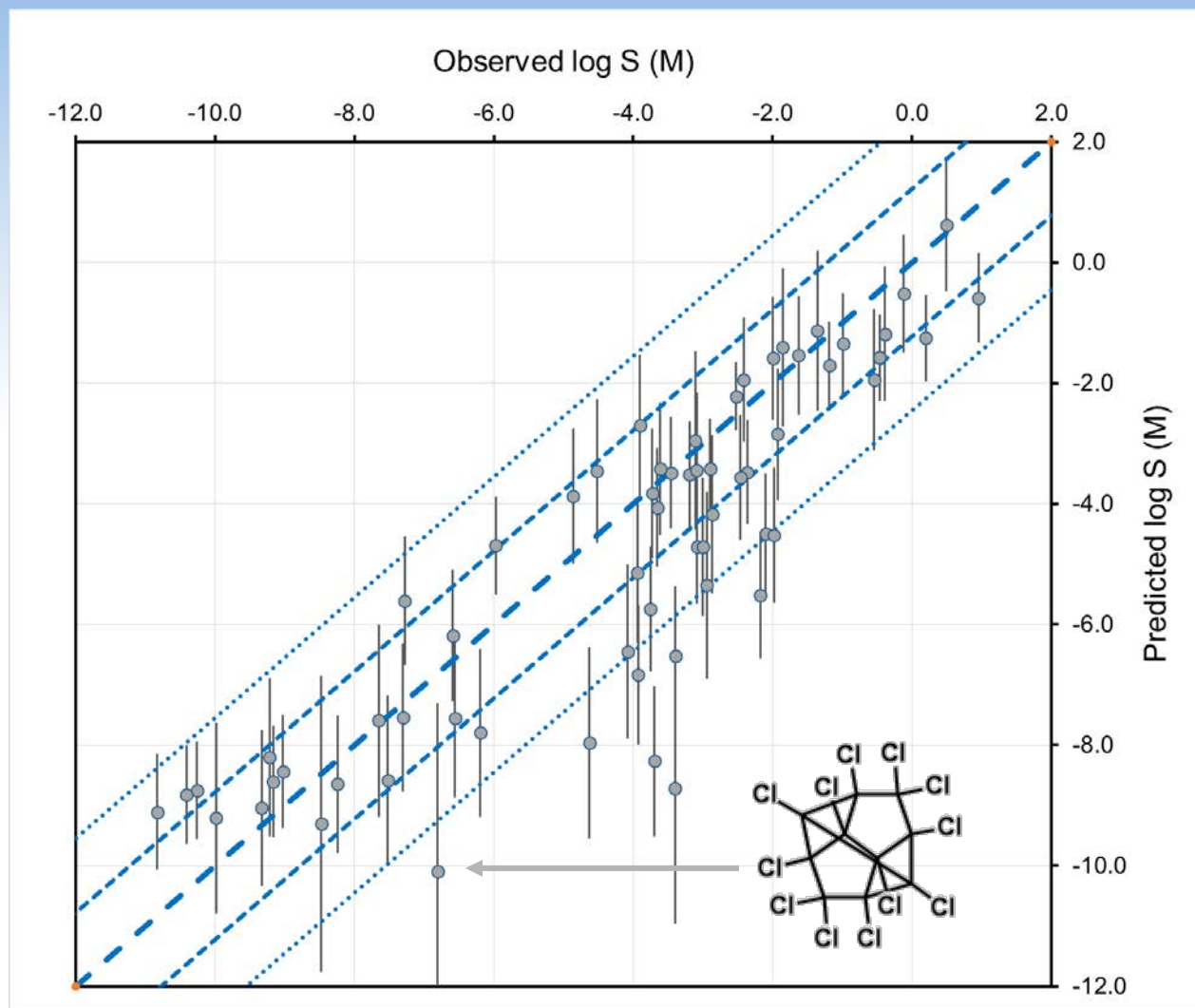
Expected distribution:
 68 : 27 : 5%

Obs. Distribution:
 67 : 22 : 11%

Where did Mirex go?



SE Estimates for Out-of-Scope Predictions



Out-of-scope Test
(n = 63)

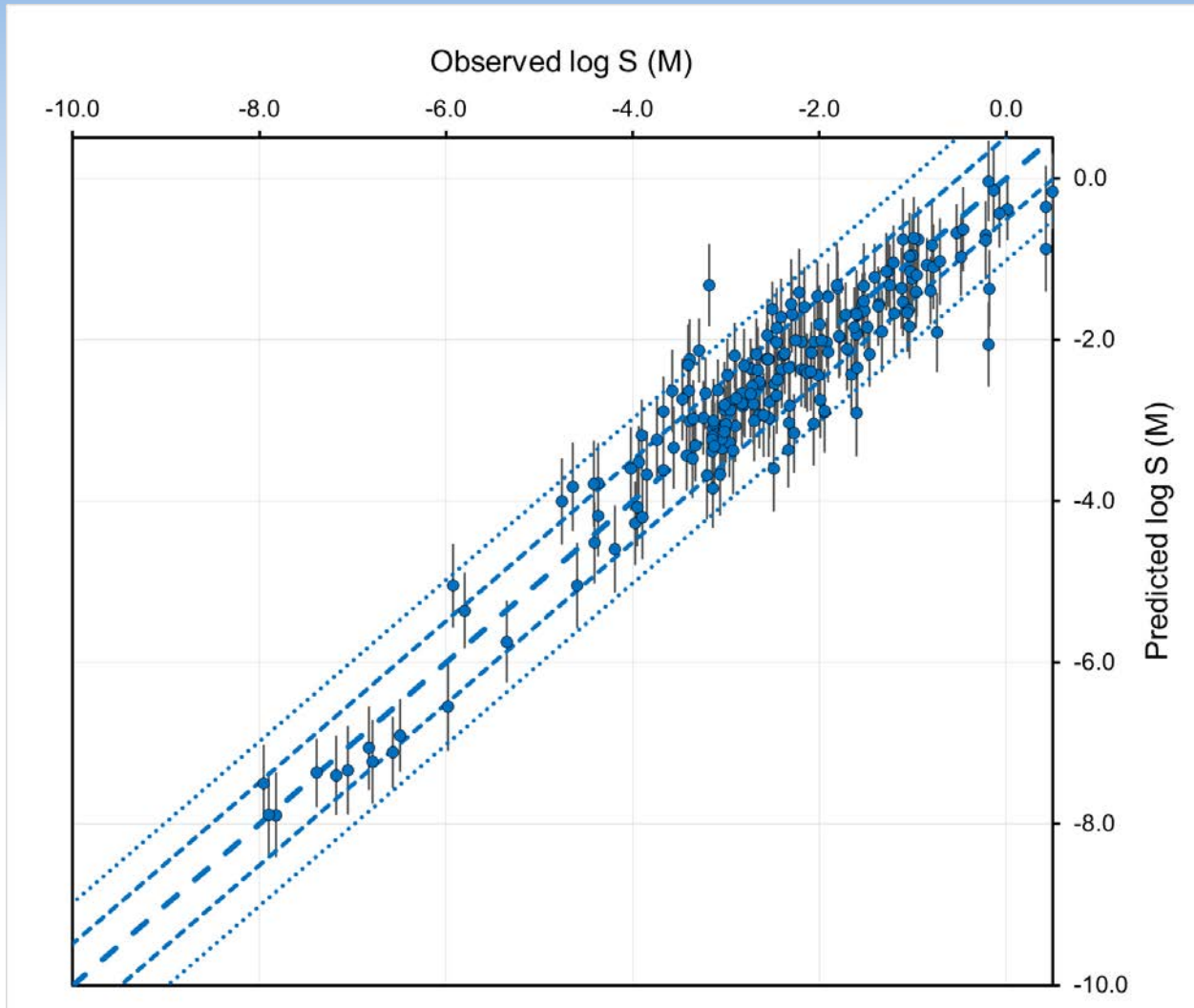
RMS SE_i (est) 1.226

RMSE (obs) 1.662

Expected distribution:
68 : 27 : 5%

Obs. Distribution:
55.5 : 28.5 : 16%

SEs Estimated for the Training Set



Best Train

$n = 291$

RMS SE_i (est) 0.509

RMSE (obs) 0.604

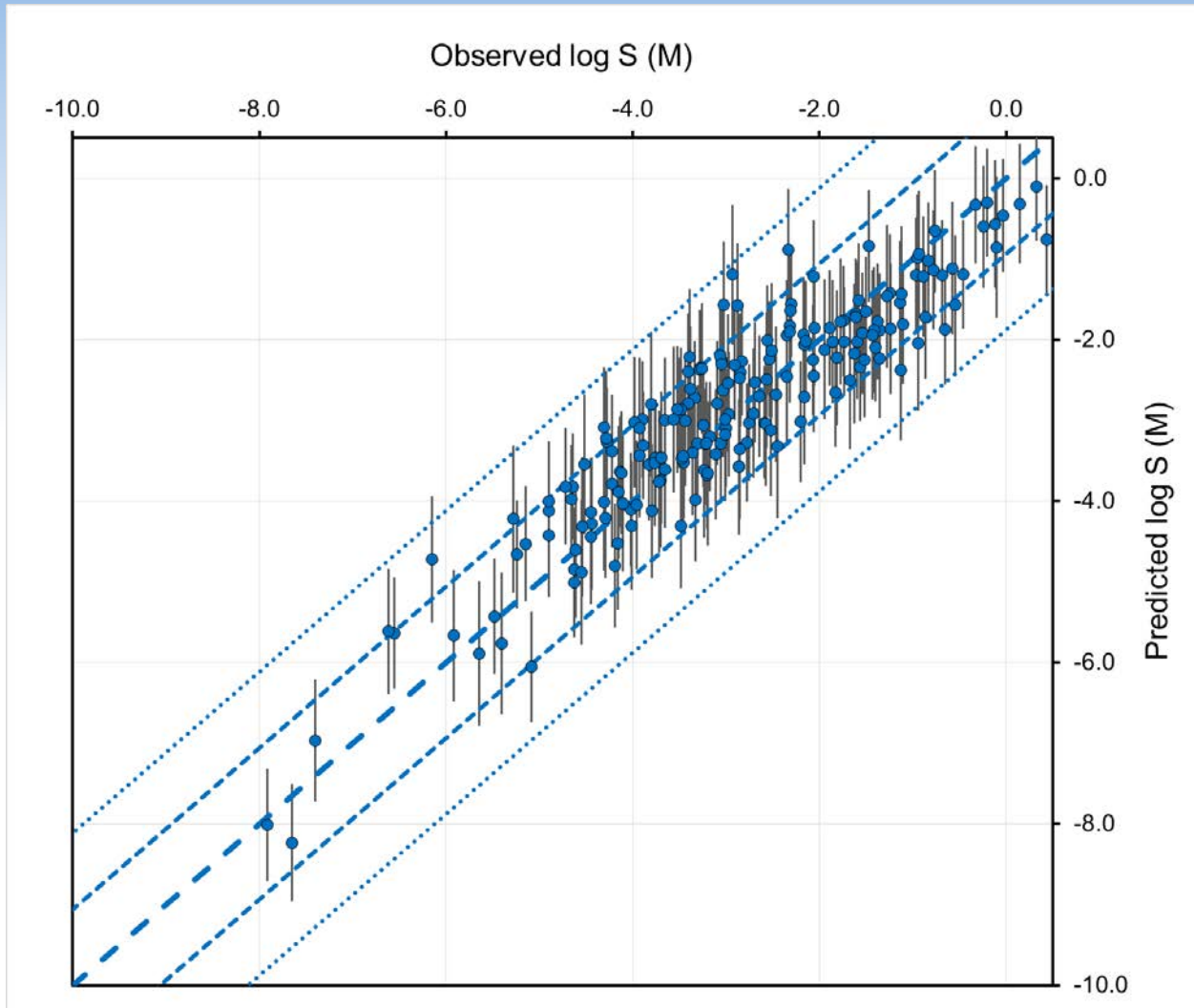
Expected distribution:

68 : 27 : 5%

Obs. Distribution:

66.5 : 26 : 7.5%

SE Estimates for the Not-So-Good Half



Less good Train
n = 292
RMS SE_i (est) 0.939
RMSE (obs) 0.729

Expected distribution:
68 : 27 : 5%

Obs. Distribution:
71 : 24.5 : 4.5%

Conclusions, Questions & Plans

- Gamma error analysis is a straight-forward approach that produces surprisingly accurate estimates of standard errors of prediction
 - it turns out to be very useful for systematically identifying bad data
 - how different can the submodels in the ensemble be?
 - how well does it work for random forest models, for example?
- Are there internal constraints that can be added to improve performance?
 - for example: α_{err} needs to be greater than α_{SD} ; early experience suggests that it should be at least 0.5 greater
- Productization and publication efforts are underway.
- We should probably come up with a better name for it...

Acknowledgements

Wenkel Liang

David Miller

Pankaj Daga

Michael Lawless

Robert Fraczekiewicz

And thank you!

bob@simulations-plus.com