Estimating Classification Uncertainty for Ensemble Models

Robert D. Clark, Wenkel Liang,
Robert Fraczkiewicz and Marvin Waldman

Simulations Plus, Inc.

Lancaster, CA 93536 USA

simulationsplus, inc.

Outline

- Motivation
- Model building in ADMET Modeler™
- Binomial, beta and beta-binomial distributions for modeling error distributions
- Fitting an uncertainty profile to the training pool
- Applying an uncertainty profile to the test set
- Using averaging instead of voting

Outline

- Motivation
- Model building in ADMET Modeler™
- Binomial, beta and beta-binomial distributions for modeling error distributions
- Fitting an uncertainty profile to the training pool
- Applying an uncertainty profile to the test set
- Using averaging instead of voting

Motivation

Drug discovery and development are a lot like poker. You cannot win consistently by being lucky. You can win consistently by knowing your opponent (Mother Nature) and by knowing the prospective odds for any given hand.

"You've got to know when to hold 'em, know when to fold 'em,
Know when to walk away, and know when to run.
You never count your money when you're sitting at the table
There'll be time enough for counting when the dealing's done."

- from "The Gambler" by Kenny Rogers

simulationsplus, inc.

More Motivation

- Drug discovery & development costs continue to rise
- Quantitative structure-activity relationships (QSARs) have the potential to speed development and reduce costs
- Regulatory agencies support the use of QSARs to guide some decisions
- Considerable progress has been made on how to accurately estimate prospective QSAR predictivity

BUT

- QSAR work has, until recently, focused on assessing the *aggregate* reliability of QSAR prediction rather than on the reliability of prospective predictions for *individual* compounds
- Researchers and regulators need to make decisions about *individual* compounds

Confidence Estimates in ADMET Predictor™ 6.5

ADMET Predictor(TM) : Ames-TRAIN-1.qmd (g:\pred uncertainty\hansen\)

File Batch Edit Calculate View Tools Help

Basic Modeler Settings | Adv. Modeler Settings | Ensemble Statistics | Model Export

Molecular Data | Prop./Desc. Histograms | Prop./Desc. Correlations | 4D Data Mining

Molecular Record Spreadsheet

Number	TOX_MUT_98	TOX_MUT_m98	TOX_MUT_100	TOX_MUT_m100	TOX_MUT_102+wp2	TOX_MUT_m102+wp2	TOX_MUT_
4393	Negative (99%)	Negative (99%)	Negative (99%)	Negative	Negative (99%)	Negative (99%)	Negative
4438	Negative (93%)	Positive (82%)	Negative (99%)	Positive	Negative (99%)	Negative (97%)	Negative
62	Negative (99%)	Positive (50%)	Negative (99%)	Negative	Negative (99%)	Negative (89%)	Negative
596	Positive (87%)	Positive (53%)	Positive (51%)	Positive	Negative (99%)	Negative (96%)	Negative

All | User Inputs | PChemBio | Metabolism | Toxicity | Simulation | Descriptors | User Models | ADMET Risk | Add Button

Current Column = 1 | 605 records | 339 descriptors | 2:16 AM

-

+

-

+

experimental Ames classification

simulationsplus, inc.

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
- L. Eriksson *et al., Environ Health Perspect* **2003**, *111*, 1361–1375
  - review of uncertainty estimation methods for QSAR
- 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

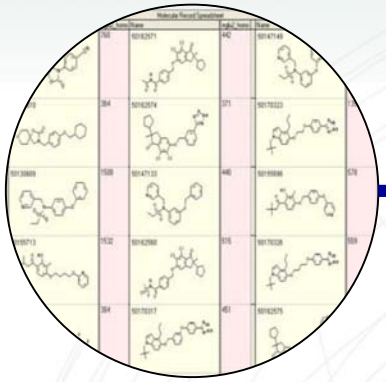
simulationsplus, inc.

Outline

- Motivation
- Model building in ADMET Modeler™
- Binomial, beta and beta-binomial distributions for modeling error distributions
- Fitting an uncertainty profile to the training pool
- Applying an uncertainty profile to the test set
- Using averaging instead of voting

simulationsplus, inc.

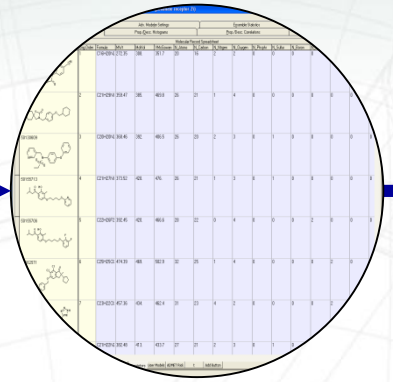
How We Build Our Ensemble Models



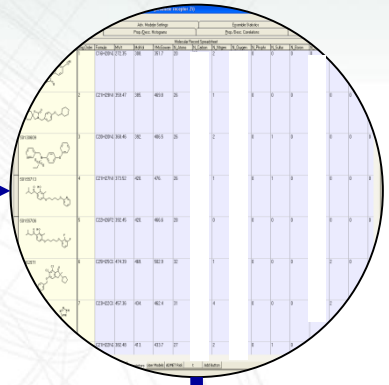
Structures and experimental data

Compute >300 Descriptors

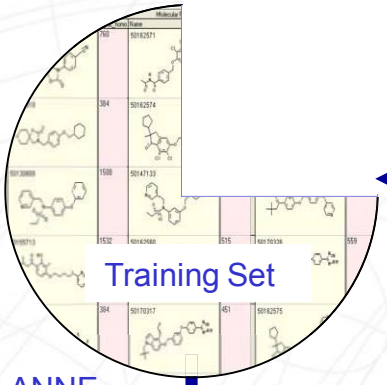
Constitutional
Topological
H bonding
Ionization
Electrotopological
Charge/reactivity



Remove low variance and correlated descriptors



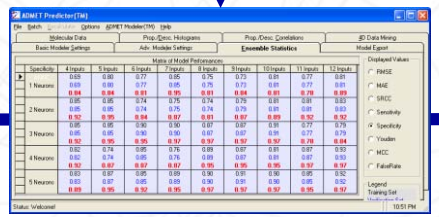
Test Set Selection
1) Kohonen map
2) Every nth
3) Random
4) K-means
5) Manual



Training Set

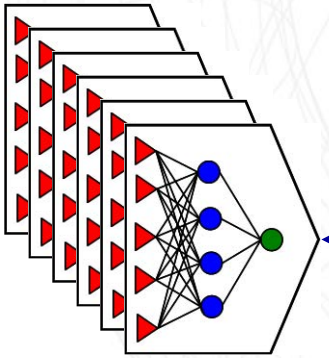
- No. of neurons and descriptors
  - Create models with different architectures
- Sensitivity analysis
  - Which descriptors create the best model

ANNE Training



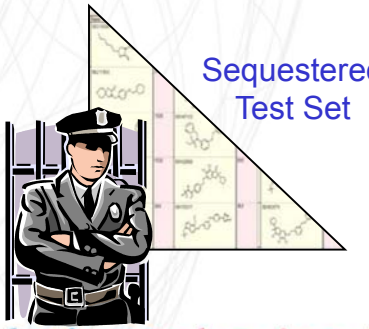
Grid of Network Ensembles
(X descriptors by Y neurons)

Select best model



ANNE

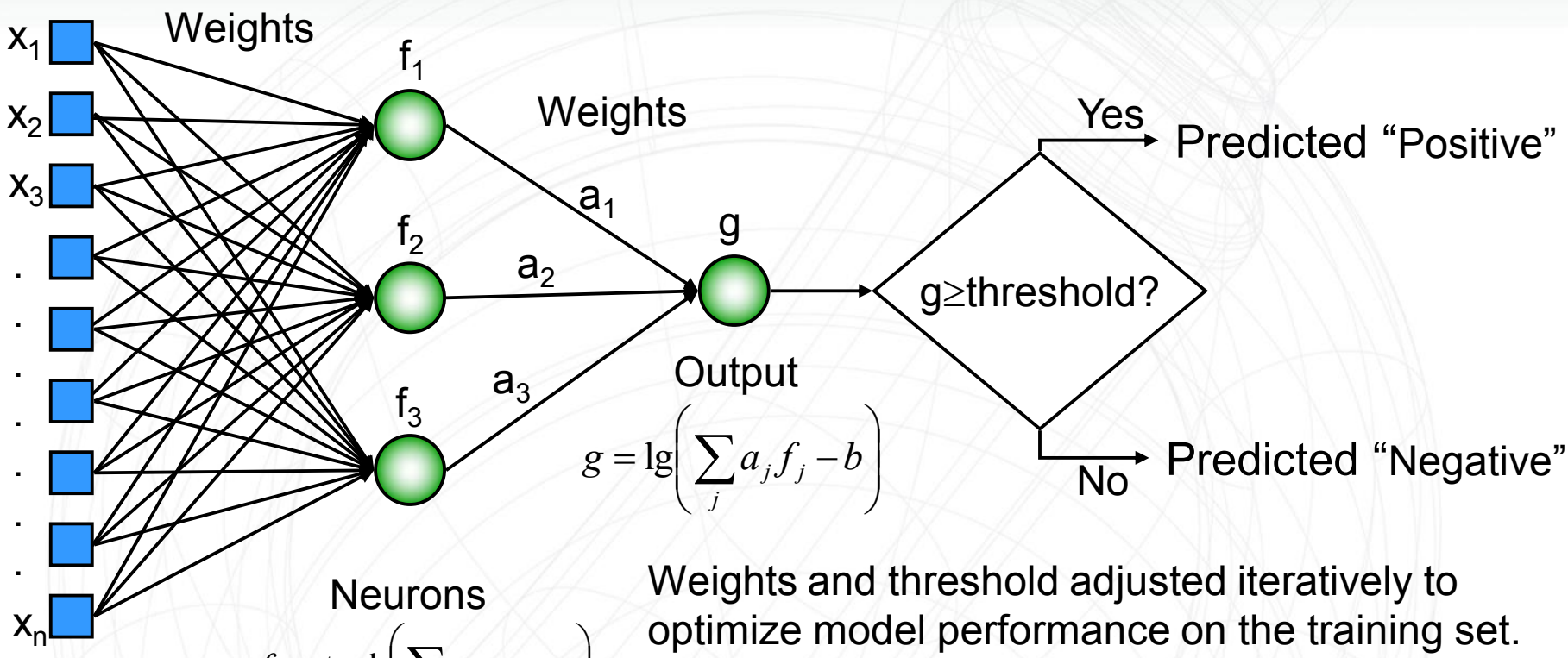
Apply model to test set



Sequestered Test Set

simulationsplus, inc.

Classification Neural Network



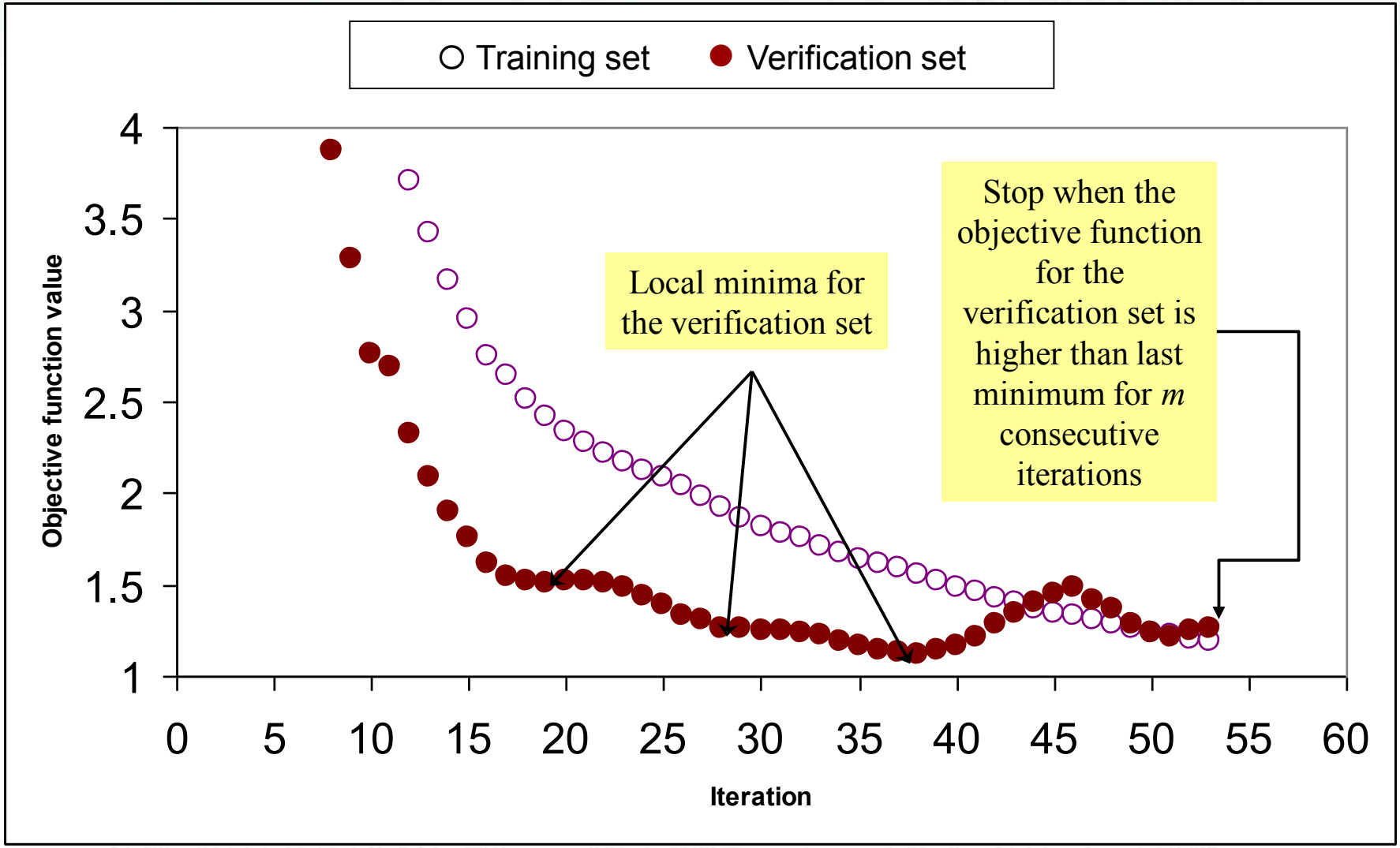
Descriptors: X_i
Normalized to range 0.0-1.0

Weights and threshold adjusted iteratively to optimize model performance on the training set.

$$Obj = \sum_{k=1}^{n} w_0 (1 - c(k))(g(k))^2 + w_1 c(k)(1 - g(k))^2$$

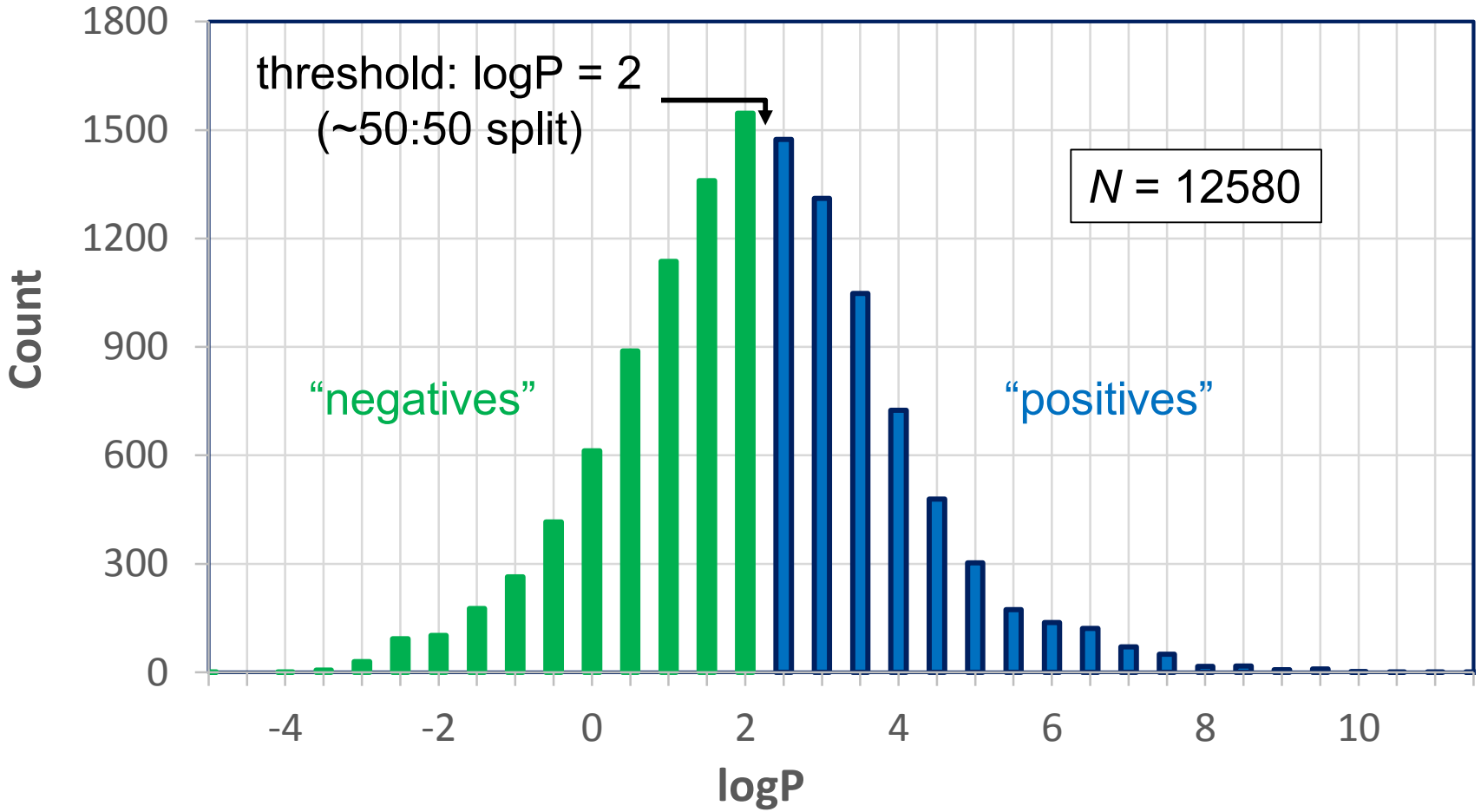
where *c*(*k*) is 0 if observation *k* is in the negative class and 1 if observation *k* is in the positive class.

Stopping Early to Avoid Overtraining



simulationsplus, inc.

The logP Data Set



simulationsplus, inc.

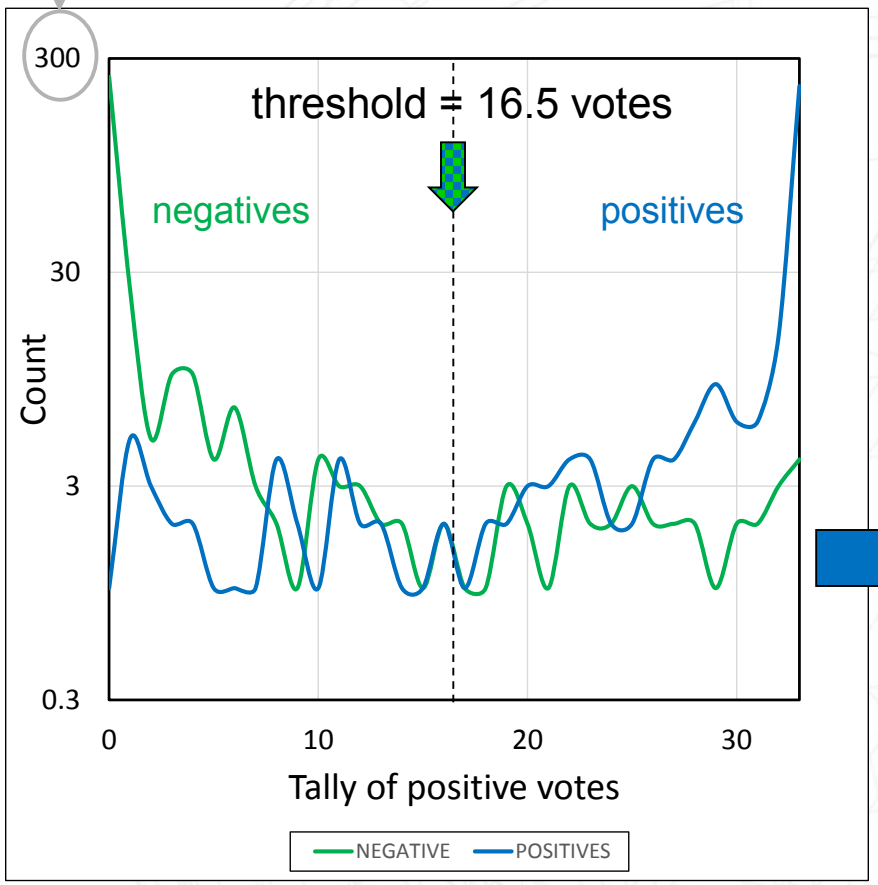
Outline

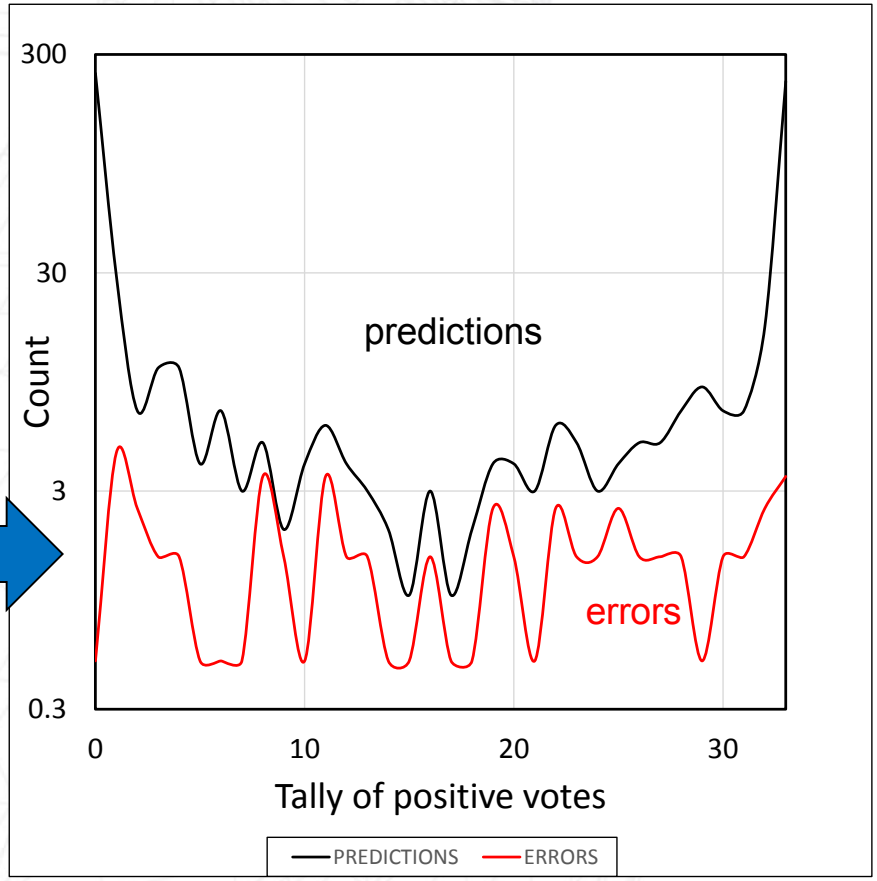
- Motivation
- Model building in ADMET Modeler™
- Binomial, beta and beta-binomial distributions for modeling error distributions
- Fitting an uncertainty profile to the training pool
- Applying an uncertainty profile to the test set
- Using averaging instead of voting

A Shift in Model Perspective

note
log scale

ANNE architecture: 36 inputs x 5 neurons x 33 networks
629 compound random training pool (train + verify)





Negatives & Positives

Predictions & Errors

simulationsplus, inc.

The Binomial Approach

- If the K network predictions are independent of one another, the errors should follow a binomial distribution across the number of positive votes k:

$$Binom(k|K,p) = \binom{K}{k} p^k (1-p)^{K-k}$$

- That grossly underestimates the spread in errors, because the networks in an ANN ensemble are not independent.
  - in addition, if they were independent the overall error rate would be expected to go down as the square root of the number of networks in the ensemble; that does not generally happen
- Tried estimating an effective number of degrees of freedom
  - that did not work very well either
- What's the alternative?

simulationsplus, inc.

Enter the Beta Binomial

- The beta binomial is a variant of the "usual" binomial distribution in which the probability of success *p* varies:

$$\mathrm{p} \sim \mathrm{B}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

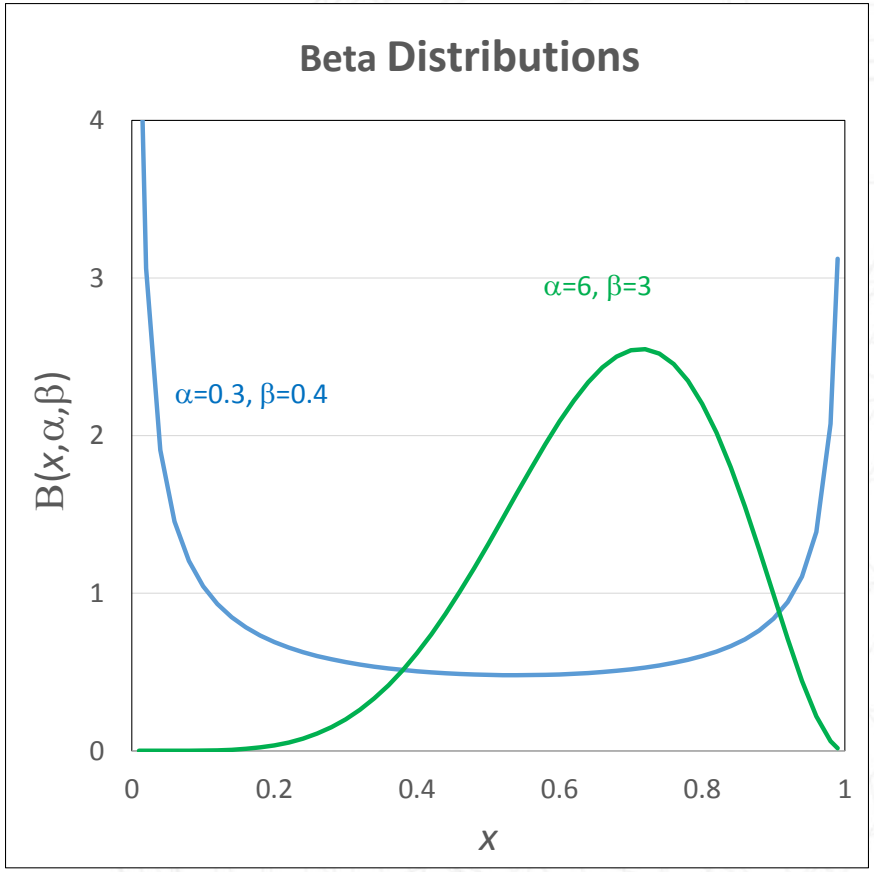
(note: $\Gamma(n) = (n\text{-}1)!$ and $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$)

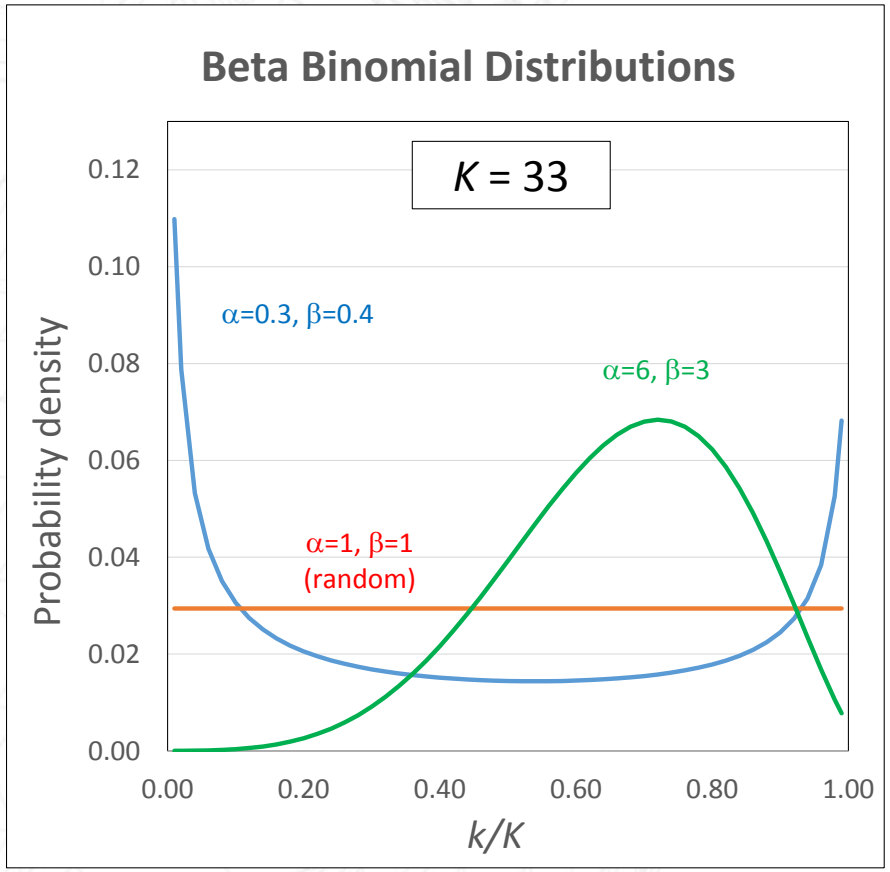
- It is used in the biometrics literature for series of events which are not independent of each other (e.g., accumulated mutations)
  - Lindsey. *Biometrics* **1999**, *55*, 449-155.
  - Dávila *et al*. *Revista Colombiana de Estadística* **2012**, *35*, 255-270

$$\mathrm{BB}(k|K, \alpha, \beta) = \binom{K}{k} \frac{\mathrm{B}(k+\alpha, K-k+\beta)}{\mathrm{B}(\alpha, \beta)}$$

simulationsplus, inc.

Meet the Beta Distributions





simulationsplus, inc.

Outline

- Motivation
- Model building in ADMET Modeler™
- Binomial, beta and beta-binomial distributions for modeling error distributions
- Fitting an uncertainty profile to the training pool
- Applying an uncertainty profile to the test set
- Using averaging instead of voting

simulationsplus, inc.

Fitting Training Pool Uncertainty

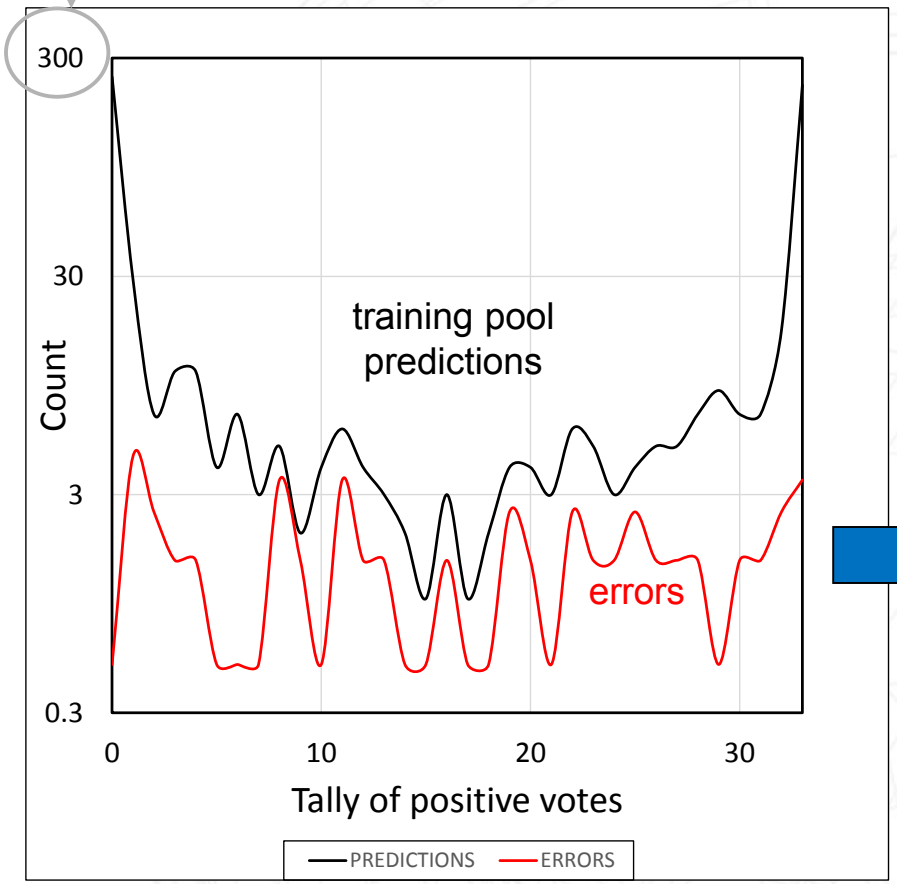
1. Build an ensemble of K networks, each with its own threshold
   - maximizing the Youden index J = (sensitivity + specificity) - 1
2. Tally the number of positive votes k for each prediction
3. Set the voting threshold to $k^* = 0.5\ K$
4. Classify negatives with $k > k^*$ as errors
5. Classify positives with $k < k^*$ as errors
6. Add a continuity correction to the count for each tally
   - necessary to mitigate problems with undersampling
   - add 1 for predictions and 0.5 for errors
7. Fit the prediction distribution to a beta binomial $\varphi(k)$
8. Fit the error distribution to a beta binomial $\varepsilon(k)$

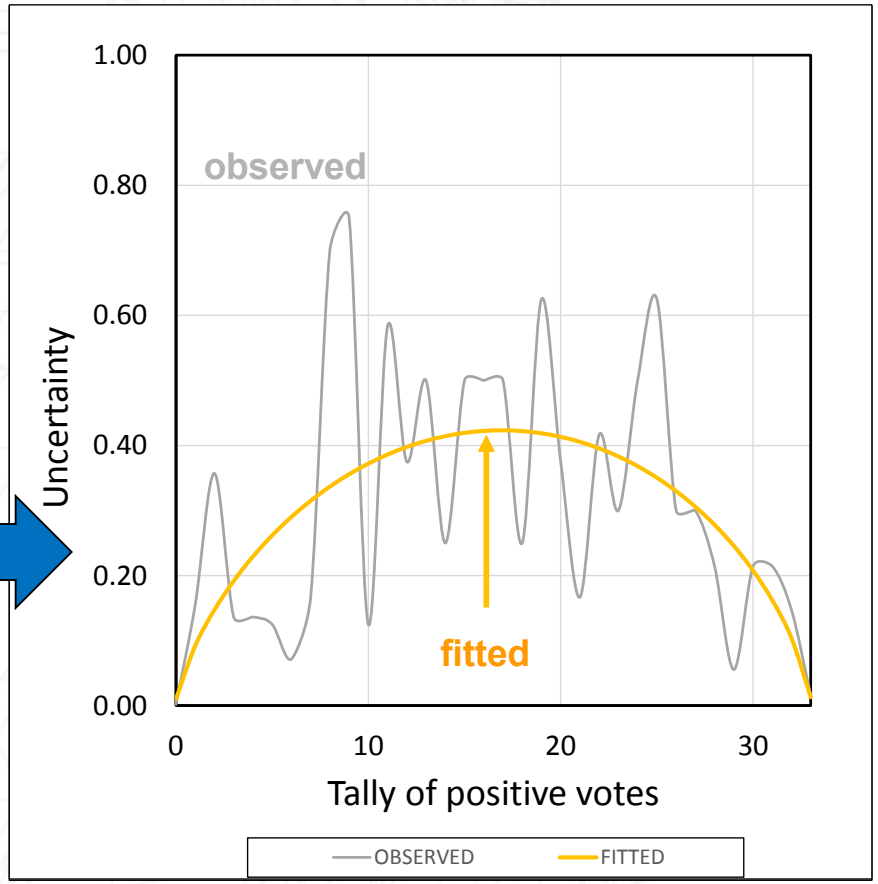
   } fit to cumulative distributions
9. Estimate the uncertainty distribution by $u(k) = FP^*\varepsilon(k)/\varphi(k)$, where FP is the overall false positive rate for the training pool
10. Calculate the estimated confidence as 1- u(k)

Training Pool Uncertainty

note
log scale

ANNE architecture: 36 inputs x 5 neurons x 33 networks
629-compound random training pool (train + verify)





simulationsplus, inc.

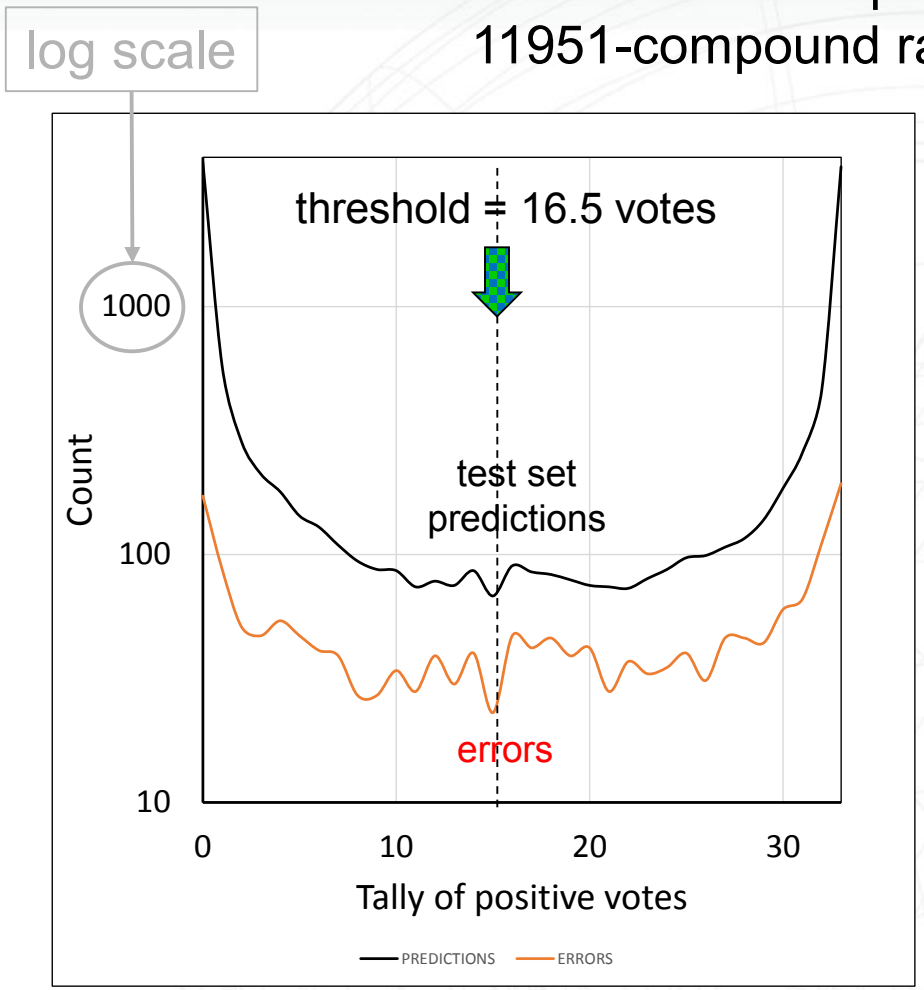
Outline

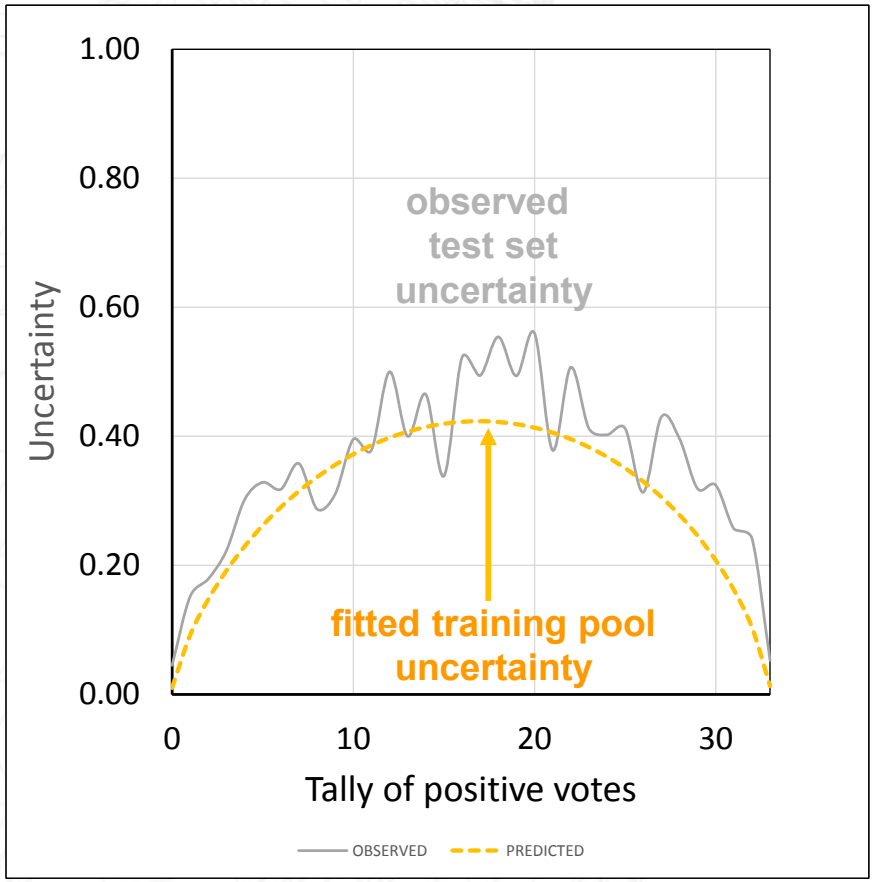
- Motivation
- Model building in ADMET Modeler™
- Binomial, beta and beta-binomial distributions for modeling error distributions
- Fitting an uncertainty profile to the training pool
- Applying an uncertainty profile to the test set
- Using averaging instead of voting

simulationsplus, inc.

It May Look Ugly, But Is It Predictive?

ANNE architecture: 36 inputs x 5 neurons x 33 networks
11951-compound random test set (95%)





YES!

simulationsplus, inc.

Why Does It Work?

- Continuity corrections suppress noise due to sparse sampling in the center of the distribution and force a limiting uncertainty of 0.5, which is the expected optimal value at the threshold

- Fitting to the cumulative distribution functions ensures that the high and low ends of the tally range – which are typically well-populated – dominate the curvatures

simulationsplus, inc.

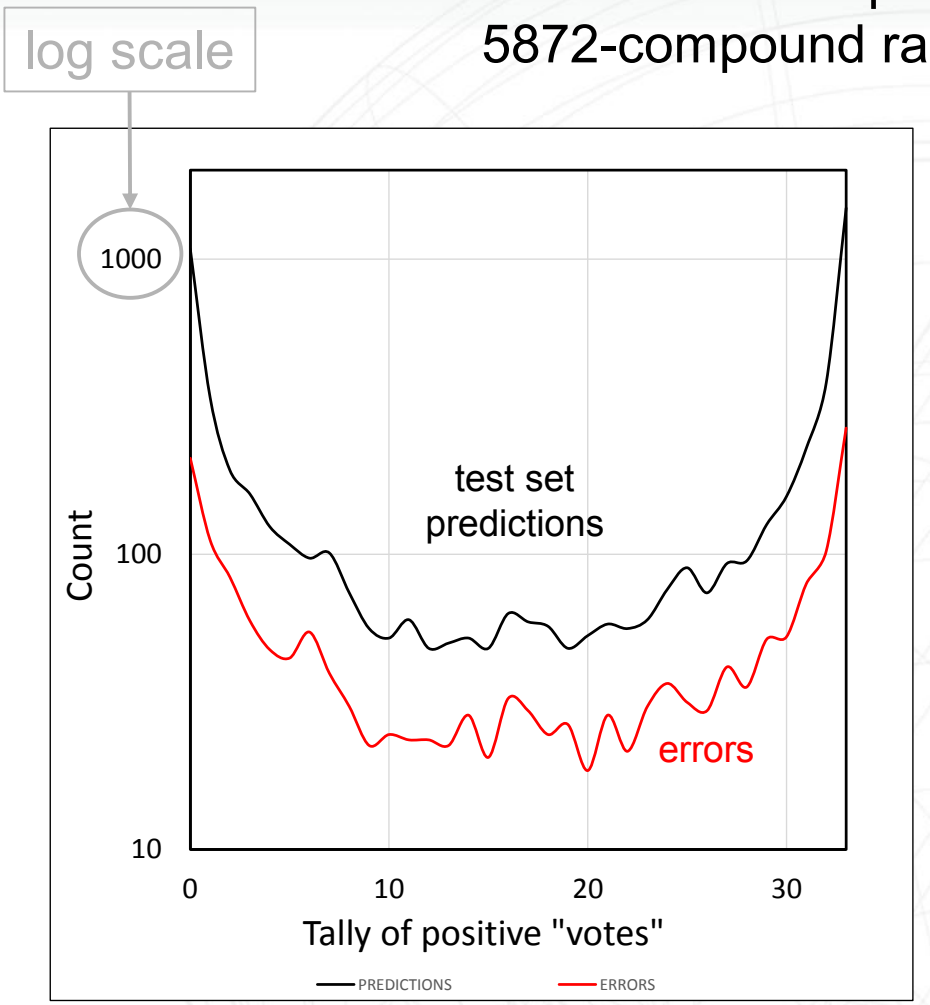
Other Examples

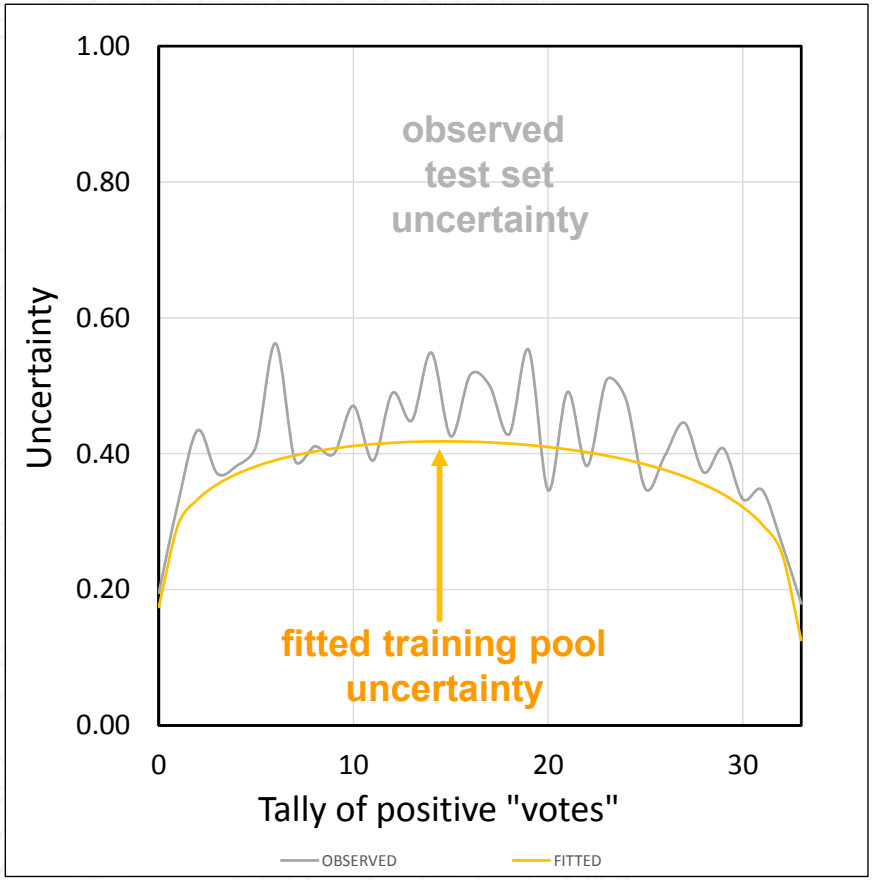
- Ames mutagenicity
  - K. Hansen *et al.* Benchmark Data Set for *in silico* Prediction of Ames Mutagenicity. *J Chem Inf Model* **2009**, *49*, 2077-2081
  - 6471 compounds used with some curation of structures
  - 2983 compounds classed as "active"
- CYP2D6 inhibition
  - NCGC luciferase-based qHTS screen: PubChem: AID 1851
  - data on 5959 compounds used with some curation of structures
  - 2806 compounds with AC50 < 10μM classed as "positive"

simulationsplus, inc.

Ames Mutagenicity

ANNE architecture: 16 inputs x 3 neurons x 33 networks
5872-compound random test set (90%)

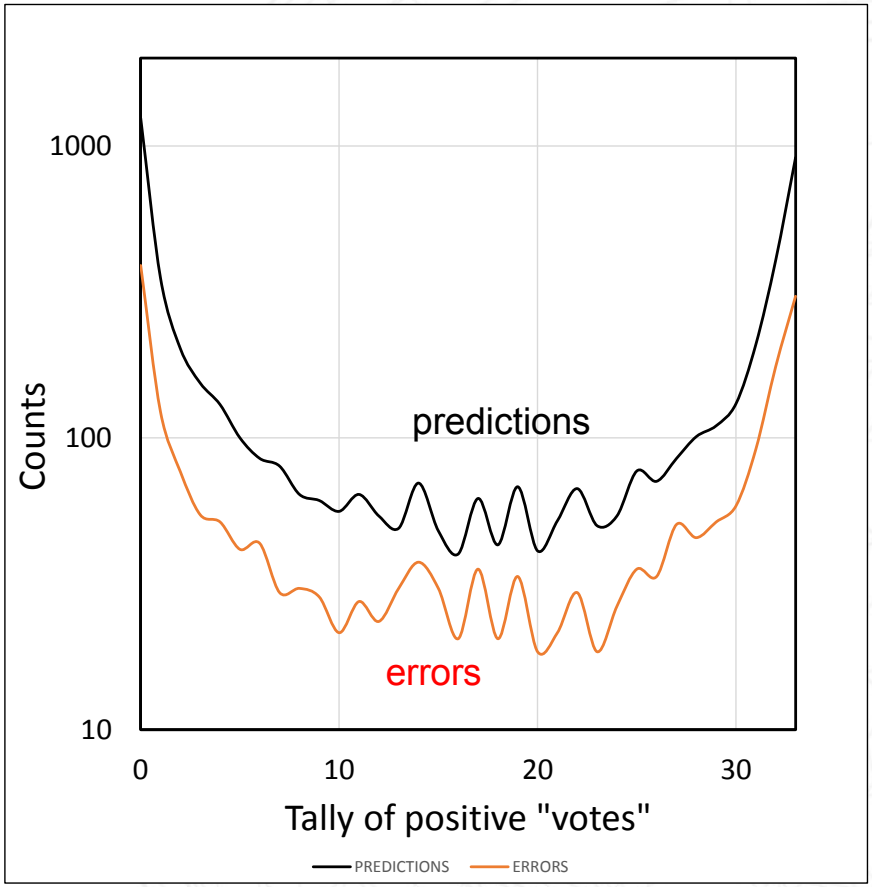


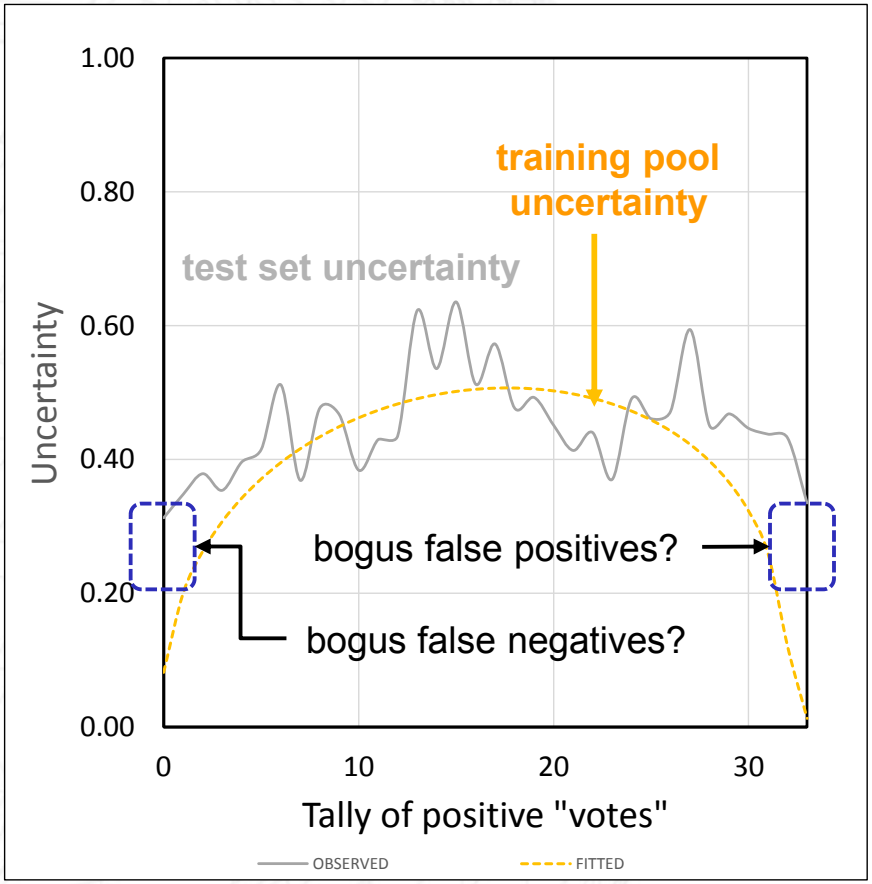


simulationsplus, inc.

CYP2D6 Inhibition

ANNE architecture: 25 inputs x 3 neurons x 33 networks
5359-compound random test set (90%)





simulationsplus, inc.

Outline

- Motivation
- Model building in ADMET Modeler™
- Binomial, beta and beta-binomial distributions for modeling error distributions
- Fitting an uncertainty profile to the training pool
- Applying an uncertainty profile to the test set
- Using averaging instead of voting

simulationsplus, inc.

Application to Averaged Outputs

Process parallels that for the voting method except that:

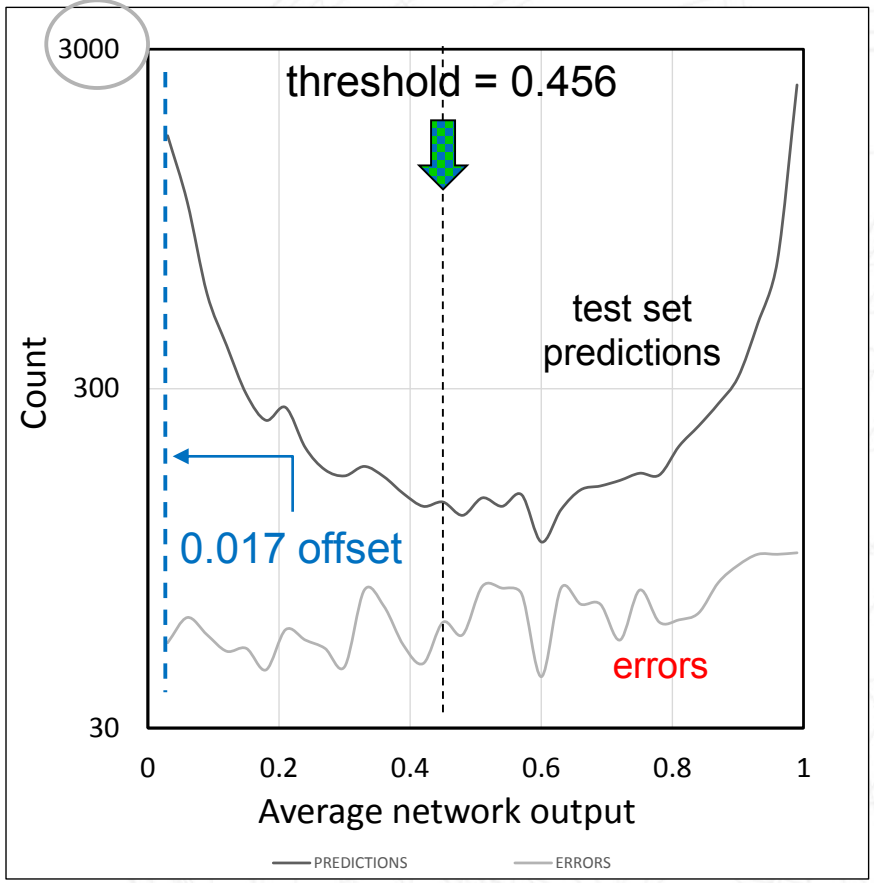
- Outputs are averaged across the networks in the ensemble and the average outputs x replace vote tallies
  - average output is a real number between 0 and 1
- The initial threshold is set to the value that provides the highest Youden index
- Predictions are collapsed inwards before fitting so as to remove "tails" at the high and low ends of the range
- The errors and prediction distributions are fit to beta functions rather than to beta binomials

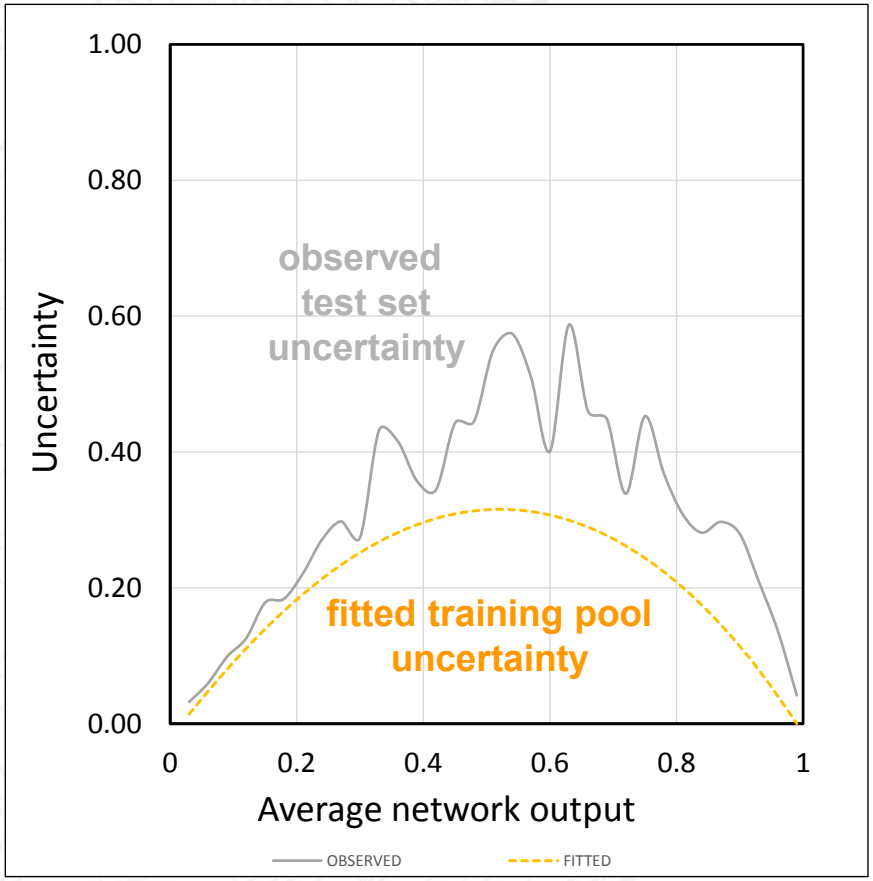
simulationsplus, inc.

logP Classification (Averaging)

log scale

ANNE architecture: 15 inputs x 3 neurons x 33 networks
11951-compound random test set (95%)





Development is ongoing…

simulationsplus, inc.

Take-Home Messages

- Fitting ensemble misclassifications to a binomial distribution across vote tallies is unlikely to work well
- ANNE prediction and error profiles follow beta binomial distributions
- The uncertainty of a prospective classification can be estimated from the results for the training pool
- Prospective uncertainty estimates are reliable for ANNEs built using early stopping to avoid overfitting
- The method used for ensemble voting can be applied to ensemble averaging by fitting to a beta distribution instead of to a beta binomial

Confidence Estimates in ADMET Predictor™ 6.5

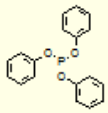
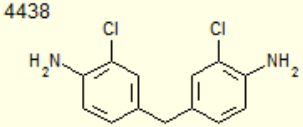
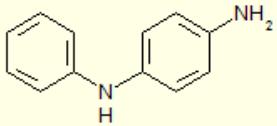
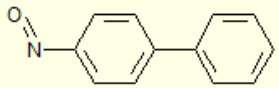
ADMET Predictor(TM) : Ames-TRAIN-1.qmd (g:\pred uncertainty\hansen\)

File Batch Edit Calculate View Tools Help

Basic Modeler Settings | Adv. Modeler Settings | Ensemble Statistics | Model Export

Molecular Data | Prop./Desc. Histograms | Prop./Desc. Correlations | 4D Data Mining

Molecular Record Spreadsheet

Number	TOX_MUT_98	TOX_MUT_m98	TOX_MUT_100	TOX_MUT_m100	TOX_MUT_102+wp2	TOX_MUT_m102+wp2	TOX_MUT_
4393 	Negative (99%)	Negative (99%)	Negative (99%)	Negative	Negative (99%)	Negative (99%)	Negative
4438 	Negative (93%)	Positive (82%)	Negative (99%)	Positive	Negative (99%)	Negative (97%)	Negative
62 	Negative (99%)	Positive (50%)	Negative (99%)	Negative	Negative (99%)	Negative (89%)	Negative
596 	Positive (87%)	Positive (53%)	Positive (51%)	Positive	Negative (99%)	Negative (96%)	Negative

All | User Inputs | PChemBio | Metabolism | Toxicity | Simulation | Descriptors | User Models | ADMET Risk | Add Button

Current Column = 1 | 605 records | 339 descriptors | 2:16 AM

simulationsplus, inc.

Acknowledgements

co-authors

- Wenkel Liang
- Marvin Waldman
- Robert Fraczkiewicz

- Michael Lawless
- Adam Lee
- Jinhua Zhang
- Michael Bolger
- Walter S. Woltosz

Thank you

simulationsplus, inc.