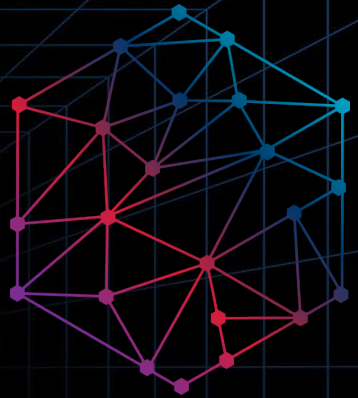
Model-Informed Drug Development

MIDD+

2021 Virtual Conference



# How We Build & Validate Industrial Strength Models

Bob Clark

SimulationsPlus

Cognigen | DILIsym Services | Lixoft

# What Does That Look Like?

A strong, reliable tool tailored to the particular demands of the relevant endpoint that will get you and your project from where you are to where you need to be safely, reliably and efficiently.

Photo courtesy of Jeffness at English Wikipedia
https://commons.wikimedia.org/wiki/File:Mackinac_Bridge.jpg

# What Makes a Good Model?

- Good data subjected to thorough curation
- Discriminating, broadly applicable descriptors
- A robust machine learning engine
- Good validation tools carefully applied
- Reliable ways to estimate predictive uncertainty

# Publication

- RD Clark & PR Daga. Building a Quantitative Structure-Property Relationship (QSPR) Model. In: *Bioinformatics and Drug Discovery,* Humana Press, New York, NY; **2019** , pp. 139-159.

# What Makes a Good Model?

- Good data subjected to good curation
- Discriminating, broadly applicable descriptors
- A robust machine learning engine
- Good validation tools carefully applied
- Reliable ways to estimate predictive uncertainty

# The Literature Is Like a River…

- The data in it changes constantly in quantity & quality
- It often contains lots of distracting things that do not really belong there
- It usually needs to be cleaned up quite a bit before you want to use what comes out of it
- The last 10% of clean-up takes at least 90% of the effort

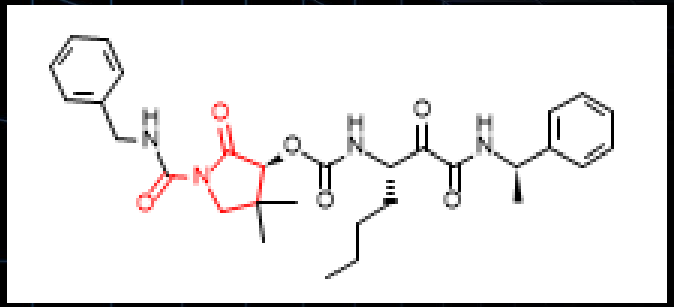
# …and Sometimes Like a Swamp

- Data comes in clumps of results for analogs that are disclosed together
  - this can and does lead to non-random errors, i.e., biases
- Data compilation can easily introduce errors but rarely removes them
- For a detailed discussion, see Phyo Phyo Zin's talk at 2:45 EST today on "Untold Stories of Data Curation"

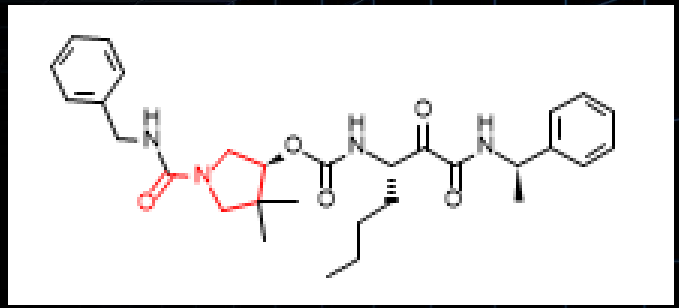
# How We Process Data

- Collect it from *diverse* sources, *uniform* endpoints
- Reconcile redundant entries ⇒ *median values*
- Standardize & analyze it graphically, then model
- Inspect outliers to find *systematic* errors
- Build more models, re-inspect outliers
- Iterate until outliers are inexplicable…or are gone

# A Recent Curation Example



8 pairs



CHEMBL382990
Bioorg. Med. Chem. Lett.
2006, 16, 1735-1739

210 nm/s

Compound 213
Curr. Top. Med. Chem.
2005, 5, 1639-1675

210 nm/s

# What Makes a Good Model?

- Good data subjected to good curation
- Discriminating, broadly applicable descriptors
- A robust machine learning engine
- Good validation tools carefully applied
- Reliable ways to estimate predictive uncertainty

# Focus on Discriminating Descriptors

- 2D molecular descriptors work well for most of the ADMET properties that we model
  - binding sites either do not exist (e.g., for $pK_a$ & solubility) or are too flexible & promiscuous for reliable docking (e.g., CYPs & UGTs)
- Simple substructure-based descriptors tend to be too localized & restrict the domain of applicability
- We focus on topological indices, electrotopological state, charge-based & ionization descriptors based on $pK_a$ analysis
- Initial choices are filtered based on degree of variance as well as both pairwise and multiway covariance

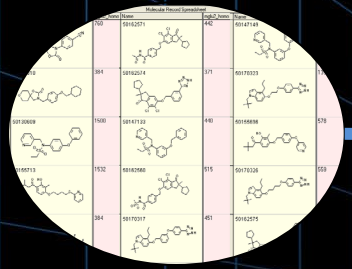
# What Makes a Good Model?

- Good data subjected to good curation
- Discriminating, broadly applicable descriptors
- A robust machine learning engine
- Good validation tools carefully applied
- Reliable ways to estimate predictive uncertainty

# Artificial Neural Network Ensembles

- ANNEs with a single hidden layer of neurons usually work well for single task ADMET learning
  - SVM, MLR and PLS are also available if needed
- Each model in the ensemble is trained on an independent random ~2:3 split of the training pool into train & verify sets
- Performance on verify sets is used to stop weight optimization before the model can become overtrained
- Multiple architectures are examined and the best performing one is kept

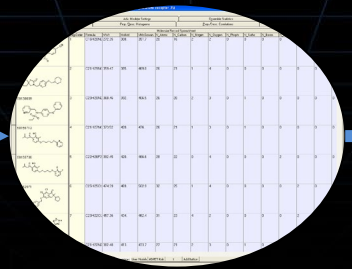
# Overview of How ANNE Models Get Built



Structures and experimental data

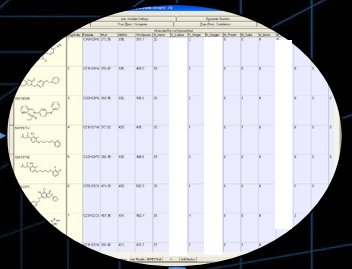
Compute

Descriptors



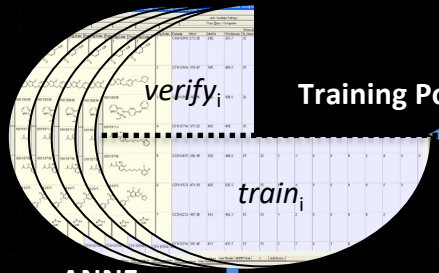
Remove low variance and

highly correlated descriptors



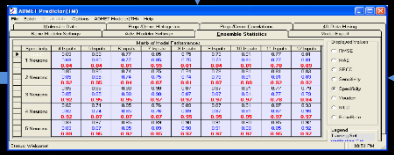
Select a Test Set

- Stratified sampling
- Kohonen map
- K-means
- Random
- Manual




Training Pool

ANNE Training

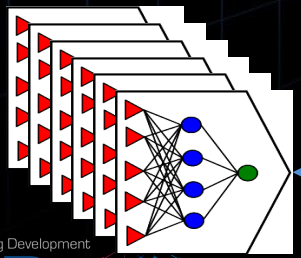


Grid of ANNE Models (*k* descriptors by *m* neurons)

Apply model to test set

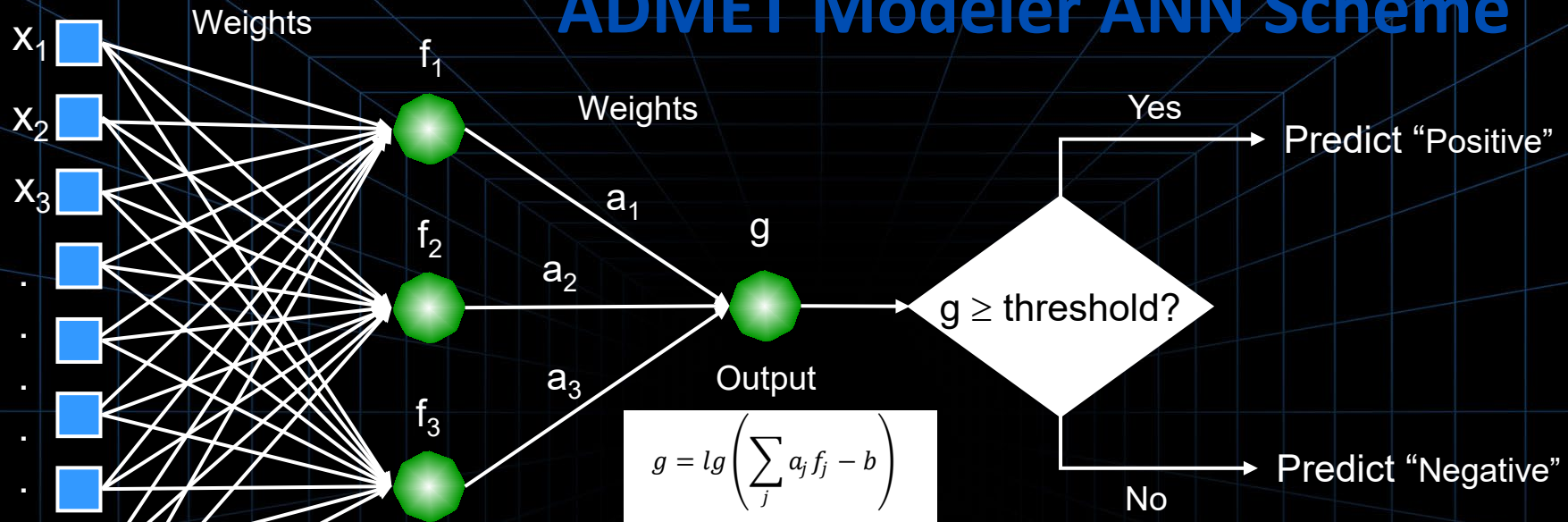
Sequestered Test Set

Select best model



Ensemble of ANNs

# ADMET Modeler ANN Scheme




$$f_i = \tanh\left(\sum_i w_{ij}\, x_i - t_j\right)$$

$$g = lg\left(\sum_j a_j f_j - b\right)$$

Descriptors $X_i$ are rescaled to lie between 0 and 1

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

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

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

# What Makes a Good Model?

- Good data subjected to good curation
- Discriminating, broadly applicable descriptors
- A robust machine learning engine
- Good validation tools carefully applied
- Reliable ways to estimate predictive uncertainty


# Estimating Predictive Uncertainty

If you stir the pile enough times and keep track carefully enough of how often your past predictions were right, you can estimate how confident you should be in the accuracy of future predictions.

# Uncertainty Methodology

- Variance in ensemble predictions is related to uncertainty, but not directly because the models are not statistically independent
  $\Rightarrow$ predictive errors follow *overdisperse distributions*
- For uncertainty in predicted classifications, we have shown that the actual errors follow a *beta binomial distribution*
  - RD Clark *et al. J Cheminformatics* **2009**, *1*, 11
- For regression uncertainty, the joint distribution of the predictive standard error and the standard deviation of prediction fit a pair of coupled *generalized gamma distributions*
  - M Waldman & RD Clark, “New approach to regression uncertainty analysis and applications to drug design,” presented Fall ACS **2019**

# Other Cheminformatics Contributors

- Marv Waldman
- Dechuan Zhuang
- Robert Fraczkiewicz
- Wenkel Liang
- David Miller
- Pankaj Daga
- Phyo Zin
- Michael Lawless
- Ola Mikosz
- Eric Jamois

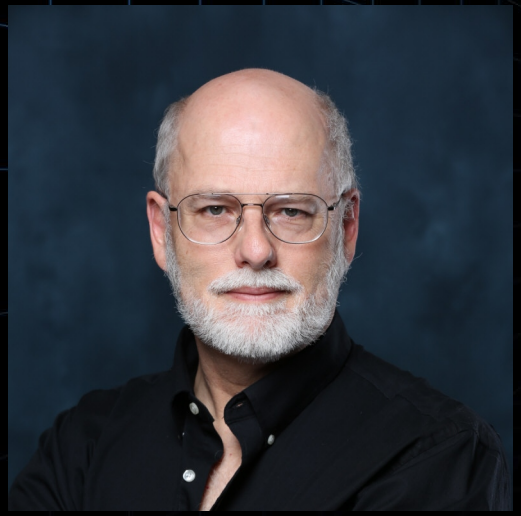
# Q&A

Questions & Answers

Model-Informed Drug Development

# MIDD+

2021 Virtual Conference



42?

*bob@simulations-plus.com*

**Learn More!** www.simulations-plus.com

SimulationsPlus

Cognigen | DILIsym Services | Lixoft