KnowItAll®
Informatics System

ValidateIt™

Statistical Validation

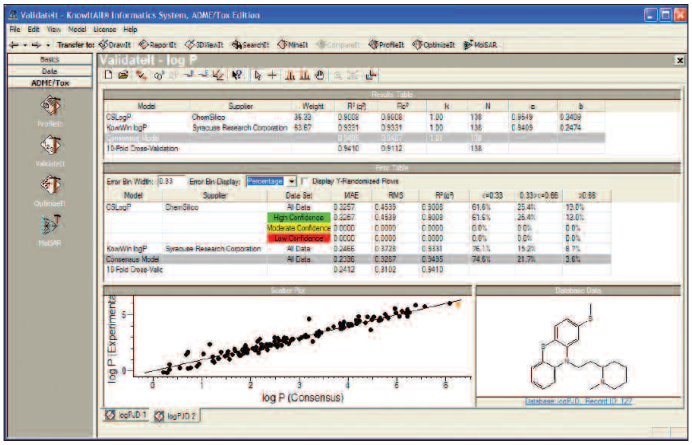
There is increasing pressure at every point on the drug discovery and development pipeline to provide documented validation of the processes used to generate results. Predicted results are no exception. Stand-alone prediction models and consensus models are easily validated by using the ValidateIt application in Bio-Rad's KnowItAll Informatics System. ValidateIt is also the tool used to generate consensus models. ValidateIt allows researchers to:

- Plot one database variable against another
- Validate the prediction of an existing model
- Create and cross-validate consensus models

Background: Models are validated by comparing the predicted results to known experimental results. Therefore, a set of known results (known set) is required for validation of any model. Once validated, the models can be used with more confidence on a larger set of compounds for which the experimental results are not known (unknown set).

In many cases, the researcher will be working on a set of proposed entities with common structural characteristics. It is recommended that the known set contain structures with characteristics similar to those in the unknown set. In this way, the validation results generated on the known set will be applicable to the unknown set.

Continuous variable models generate results in the form of a real number (for example, log P, or water solubility. Classification models generate Boolean results (for example, True or False, as in the case of mutagenicity). Both model types can be validated within KnowItAll using the ValidateIt application.



Using ValidateIt with Continuous Models: Continuous models generate results in the form of a real number. For continuous models, ValidateIt reports the following statistics:

- Weight, model weighting used in the calculation of a consensus model (not reported if only one model is used)
- R^2, Correlation coefficient between predicted and actual values in a model for the best-fit line
- q^2, Correlation coefficient between predicted and actual values in validation set
- R_0^2, Correlation coefficient between predicted and actual values in a model for the best-fit line going through the origin
- N, Number of compounds in the validation set
- k, Slope for regression through the origin (k)
- MAE, Mean Absolute Error
- RMS, Root Mean Squared Error
- Error Bins (low, med, high)
- Scatter plot of actual vs. predicted and chemical structure (selected by clicking on scatter point)

If ValidateIt is being used within a *consensus model* (COMO), then the user can view the individual statistics for each of the individual models, including a scatter plot. COMOs are created using an integrated N-fold cross-validation technique, assuring the most accurate results possible.

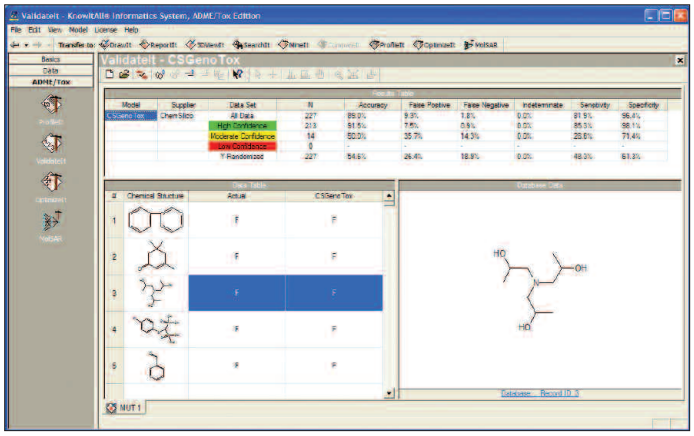
If a *localized consensus model* (LOCOMO) is being used, the conditional substructures are also shown in ValidateIt. A LOCOMO is a COMO that is trained on a subset of compounds that include and/or exclude user-specified substructures. In this way, consistent substructure-specific errors in models can be accounted for, providing significantly more accurate results than for a simple COMO.

BIO-RAD

Using ValidateIt with Boolean Models: Boolean prediction models generate True/False results and are handled differently inside ValidateIt. An example would be a mutagenicity prediction model, wherein the result of the prediction is either "Mutagenic" (True) or "Non-Mutagenic" (False). KnowItAll provides an integrated normalization function to accurately map real variable predictors to a consistent Boolean scale.

When used with Boolean models, ValidateIt reports the following information:

- Accuracy - The percentage of all predictions, True or False, that are correct
- False Positive - The percentage of True predictions that are not correct
- False Negative - The percentage of False predictions that are not correct
- Indeterminate - The percentage of items not calculated
- Sensitivity - The accuracy of True predictions
- Specificity - The accuracy of False predictions
- Data Table - Shows True/False results for each model used along with the actual experimental value



ValidateIt includes an automated cross-check that involves randomization of the experimental values, prediction, and then a comparison of the predicted versus the randomized actual results. This method is a standard technique in Boolean model validation, in which the results of predictions with actual results are compared with randomized results as a guard against a poor model giving good results simply by chance.

When multiple Boolean models are used in a consensus modeling environment, ValidateIt enables the user to choose the method by which the consensus will be determined:

- Worst Case Scenario - If any model predicts True, then the consensus will be True
- Best Case Scenario - If any model predicts False, then the consensus will be False
- Majority Rules - More than 50% of the models must predict either True for the consensus to be True (or False for the consensus to be False). If an even number of models are used, and the determination is 50% True and 50% False, then the consensus will be "Indeterminate."
- Percent Agreements - Similar to majority rules, but the user can set the Percentage Agreement among available models required to reach a consensus result

Summary: Validation of prediction is an essential step in the successful use of prediction tools in a regulated environment. The KnowItAll Informatics System, including ValidateIt, provides the tools needed to produce validation statistics specific to the active project. Conveniently, the ValidateIt application also serves as the control center for viewing and evaluating the suitability of a single model or consensus model.

BIO-RAD

Bio-Rad Laboratories, Inc.

Informatics Division
www.knowitall.com

China	*Phone: +86 010 5939 0088 x381 • Email: informatics.china@bio-rad.com*
Europe, Middle East, Africa	*Phone: +44 20 8328 2555 • Email: informatics.europe@bio-rad.com*
India	*Phone: +91 124 4029300 • Email: informatics.india@bio-rad.com*
Japan, Taiwan, Korea	*Phone: +81 3 (6361) 7080 • Email: informatics_jp@bio-rad.com*
USA	*Phone: +1 267 322 6931 • Toll Free: +1 888 5 BIO-RAD (888-524-6723) • Email: informatics.usa@bio-rad.com*
All Other Countries	*Phone: +1 267 322 6931 • Email: informatics.worldwide@bio-rad.com*

96222-REV201107