

Modeling Stressed HTS Data

Phil Cochrane, Ph.D.

pcochrane@scitegic.com



confidential

*ask **more** of your **data***

Organization of Talk

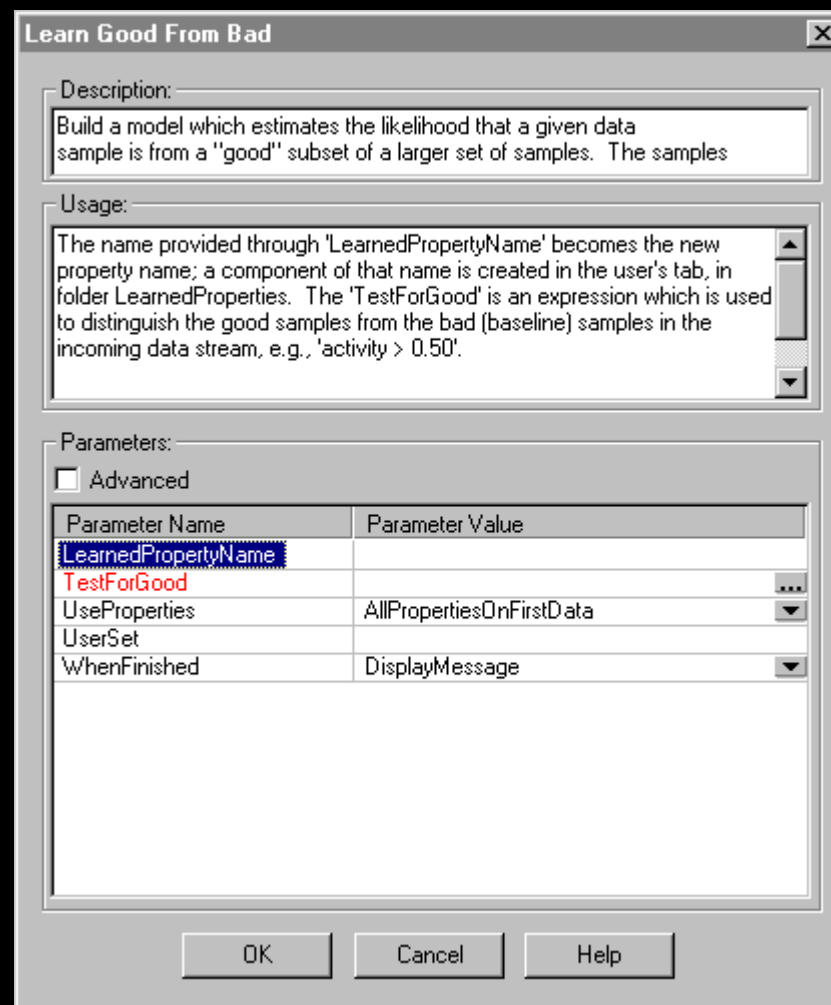
- Modeling HTS Data
 - Description of Bayesian Learning
 - The NCI AIDS HTS data set
 - Modeling results on original data set
 - Results after stressing the data

Bayesian Learning

- Build a model which estimates the likelihood that a given data sample is from a "good" subset of a larger set of samples (*classification learning*)
- SciTegic uses modified Naïve Bayesian statistics
 - Efficient:
 - scales linearly with large data sets
 - Robust:
 - works for a *few* as well as *many* 'good' examples
 - Unsupervised:
 - no tuning parameters needed
 - Multimodal:
 - can model broad classes of compounds
 - multiple modes of action represented in a single model

Learning: “Learn Good From Bad”

- Use calculated or experimental properties
 - Typical: fingerprints, logp, donors/acceptors, molecular property counts
- Test for good, e.g.:
 - Activity > 0.5
 - MA RLIKE ‘ESTRO’
 - ScreenResult = 1
- Model becomes new component in system

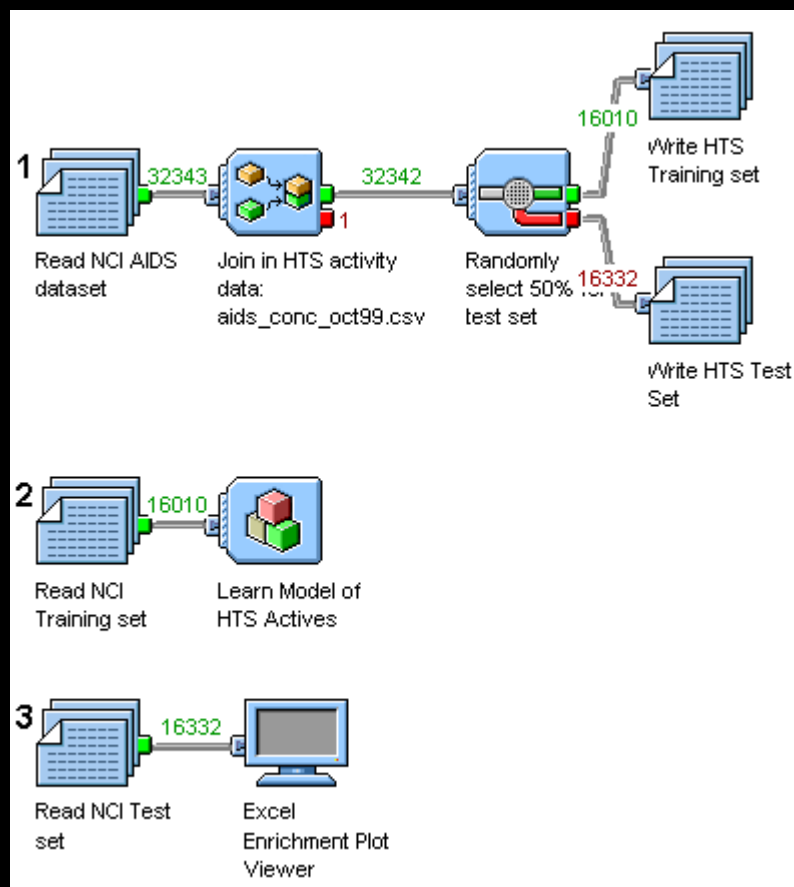


The NCI AIDS data

- Learning illustrated with NCI AIDS data set
- ~32,000 compounds selected for HTS
- Whole-cell assay
- Found 230 confirmed hits (“CA”)
- Found 230 weak hits (“CM”)
- Represent 7 “activity classes” (modes of activity)

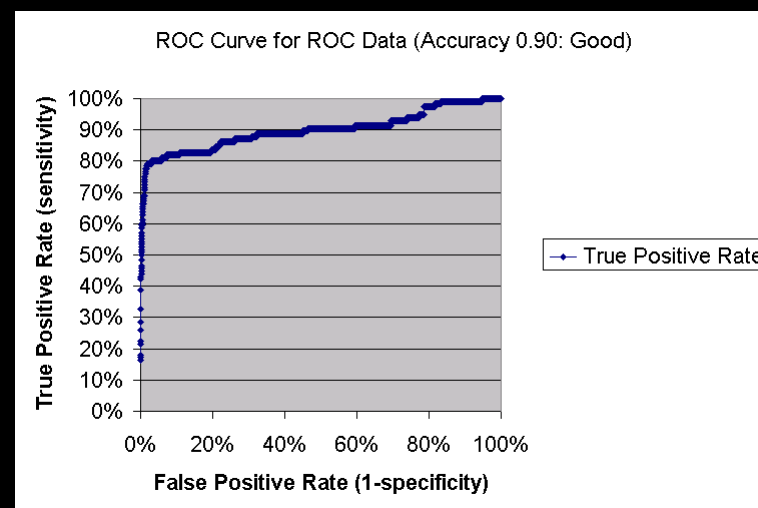
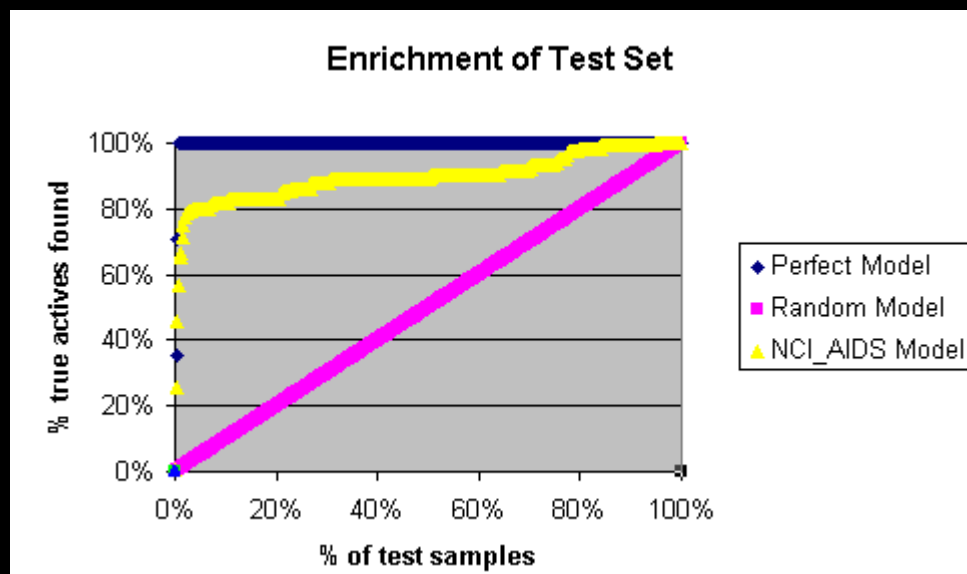
Experiment 1: Modeling the Data

- Data split 50/50
- Trained on 16,000 samples w/ 115 hits



Results

- Results:
 - Would have discovered 80% of actives screening ~600 compounds
 - Model learned *multiple modes of activity*



Modeling of stressed HTS data

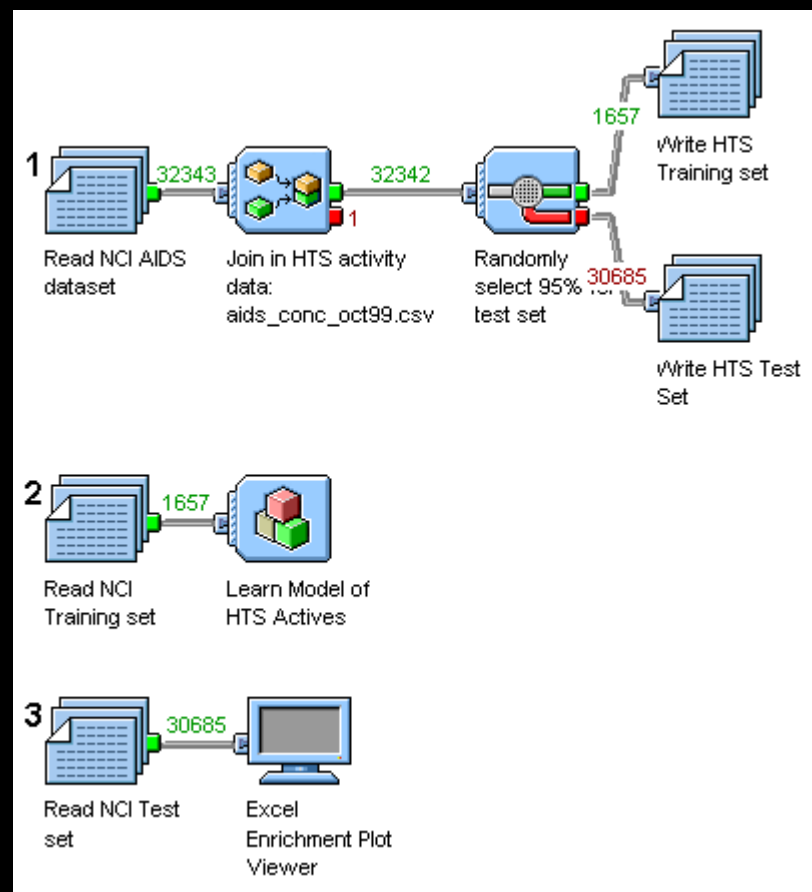
- Bayesian modeling will be applied to the results of an HTS screen after applying *stresses* to the data
- The process:
 - The raw data is *stressed*
 - The data is split into *training* and *test* sets
 - Model built using training set
 - Model applied to test set
 - Enrichment plot used to visualize effect of stress

The stresses and problems

- Multiple modes of activity (present in original data)
- Small screening sets w/ few hits
- Noisy (unconfirmed or convoluted) hits
- Poor-quality (weakly active) hits
- Identifying false positives and “singletons”
- Recapturing false negatives

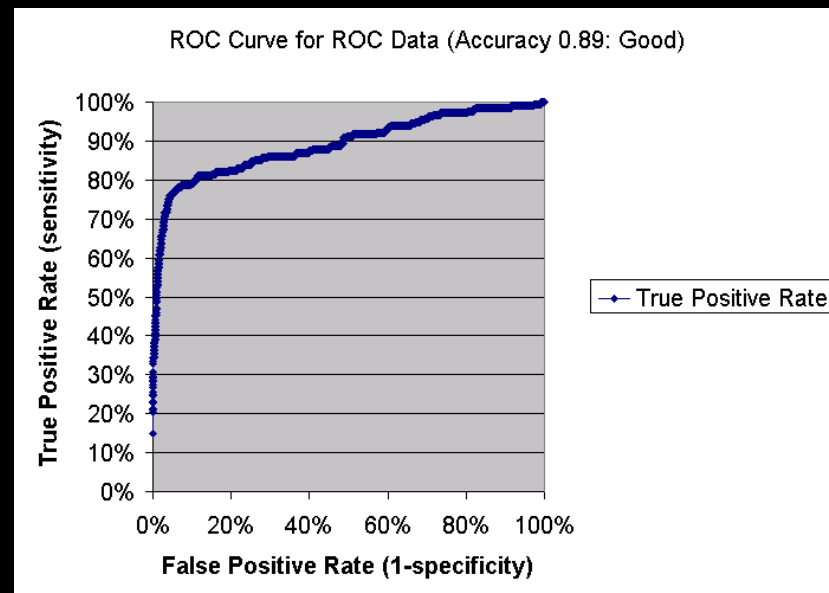
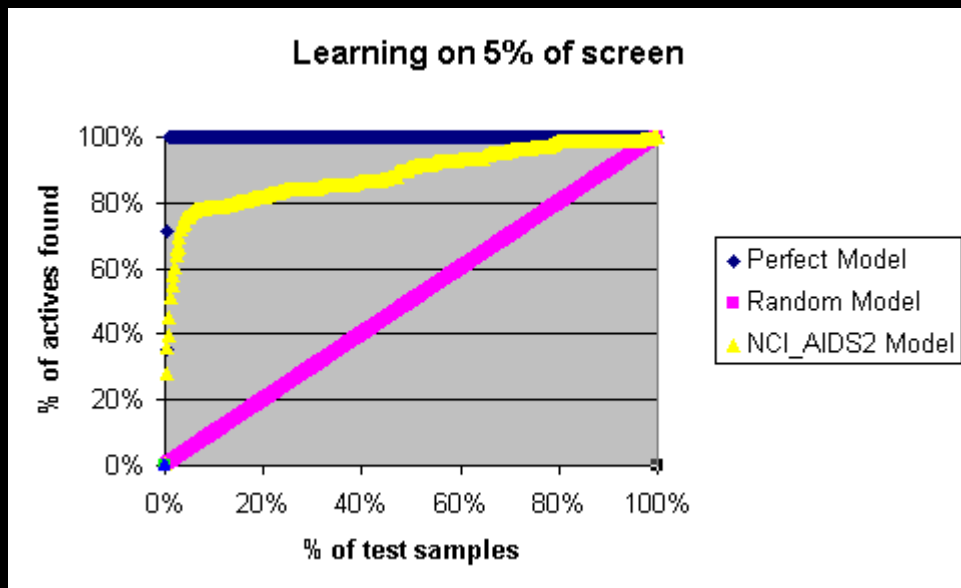
Experiment 2: Small number of hits

- Data split 5/95
- Trained on
 - ~1,600 samples
 - 14 hits



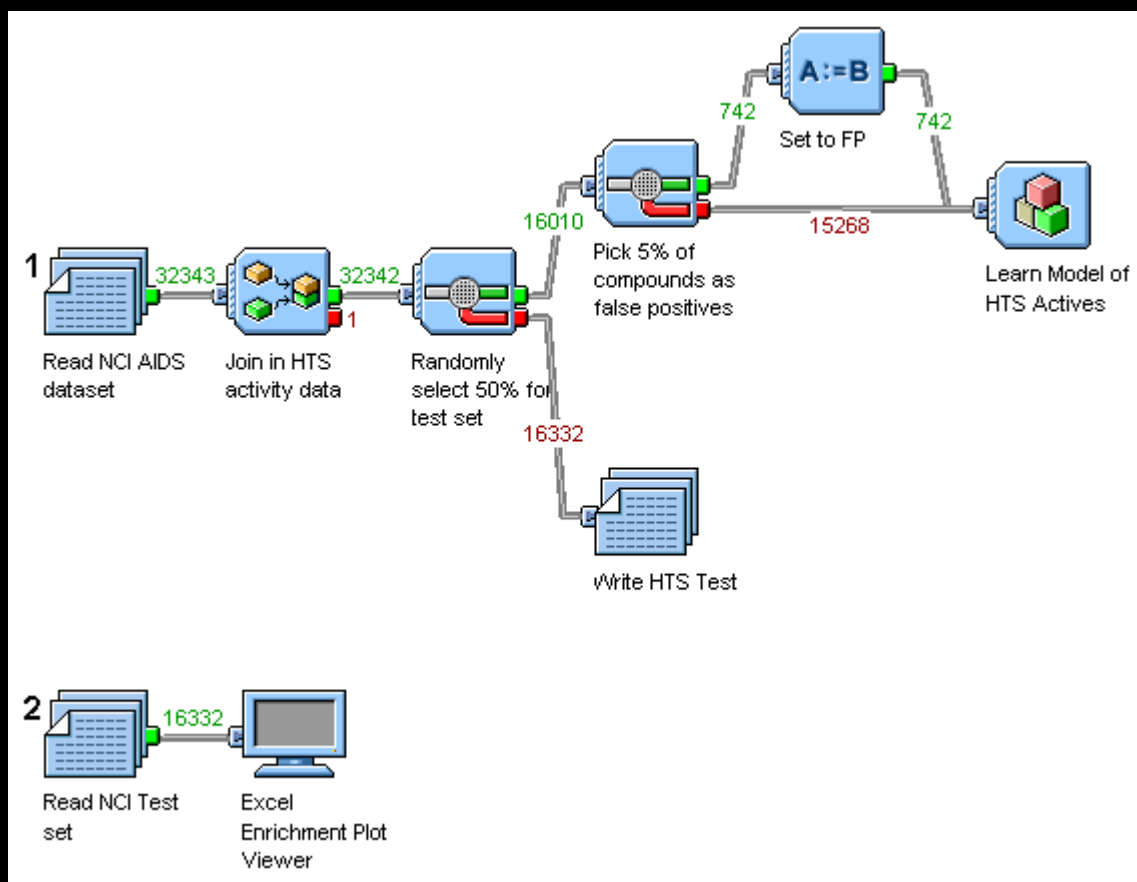
Results

- Results:
 - Would have discovered 80% of actives screening ~3,000 compounds



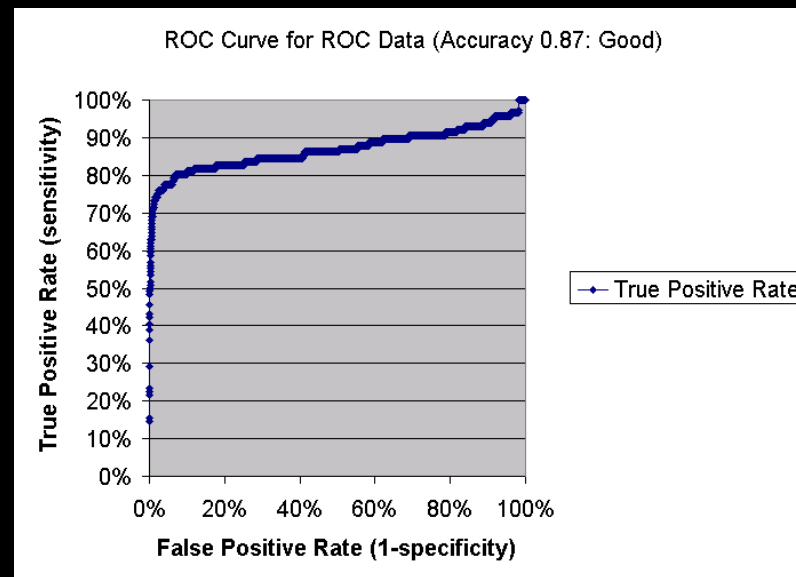
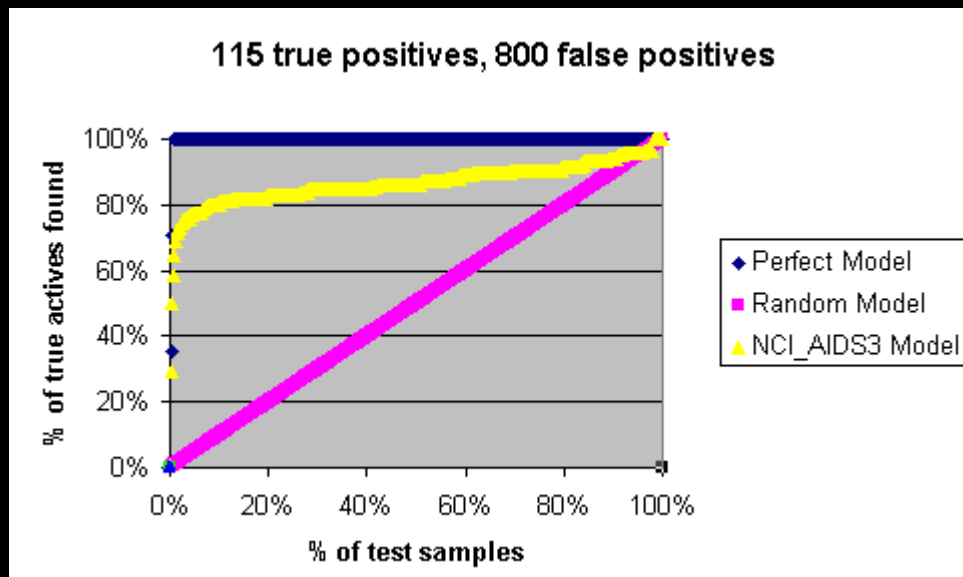
Experiment 3: Noisy hits

- Data split 50/50
- 5% of negatives in training set reassigned as *false positives*
- Final training data contained 115 *true* actives and ~800 *false* actives



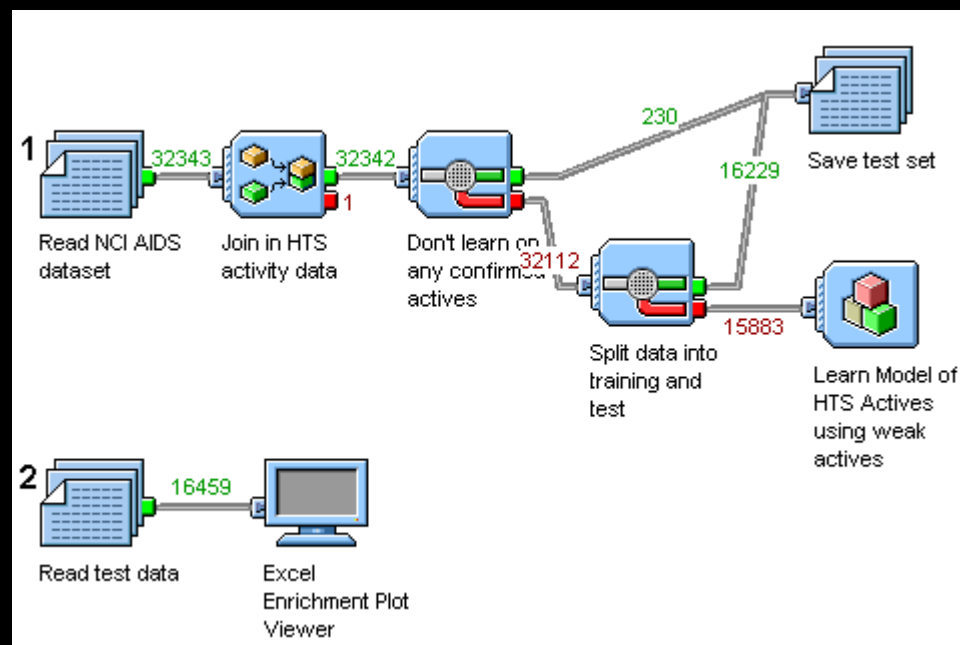
Results

- Results:
 - Would have discovered 80% of actives screening ~1,500 compounds



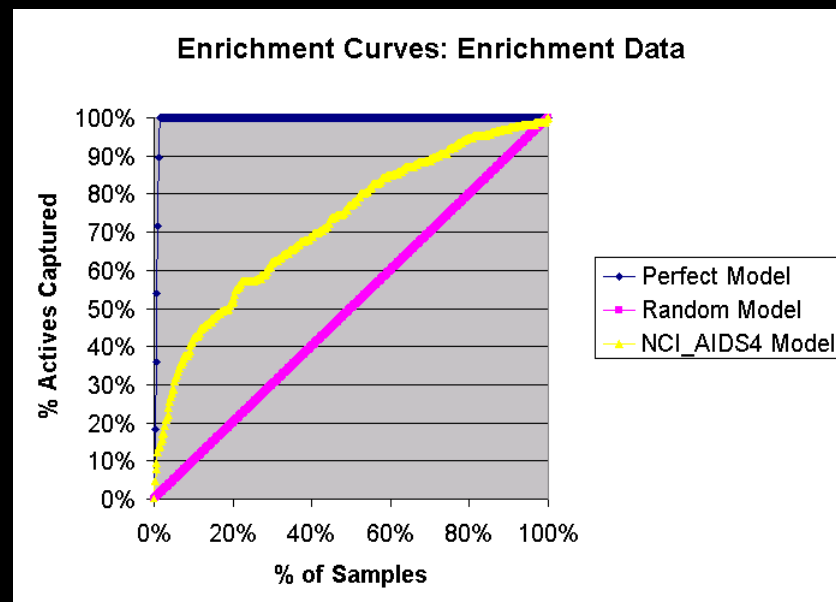
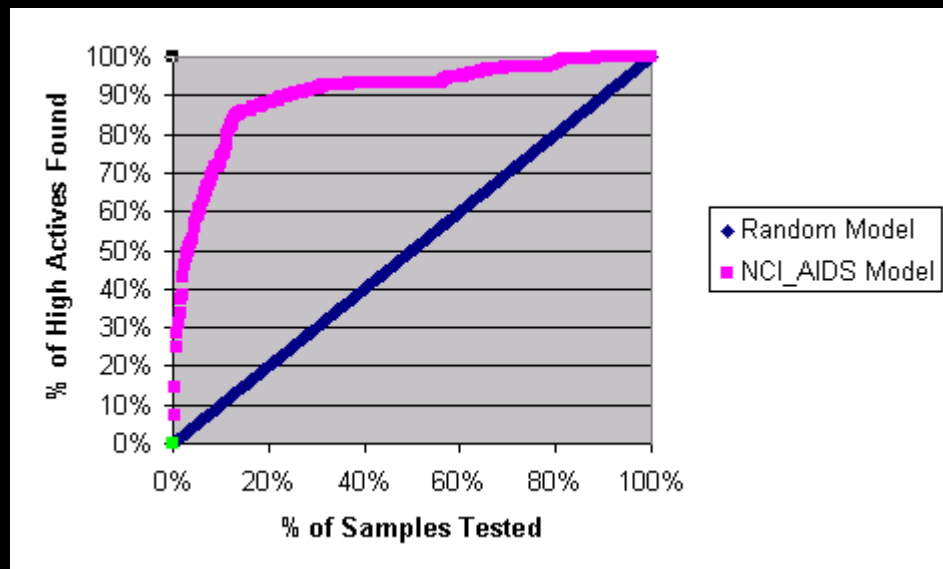
Experiment 4: Weak hits

- Data split 50/50
- All confirmed actives (CA) removed to test set
- Trained on 130 confirmed *moderately active* (CM) compounds



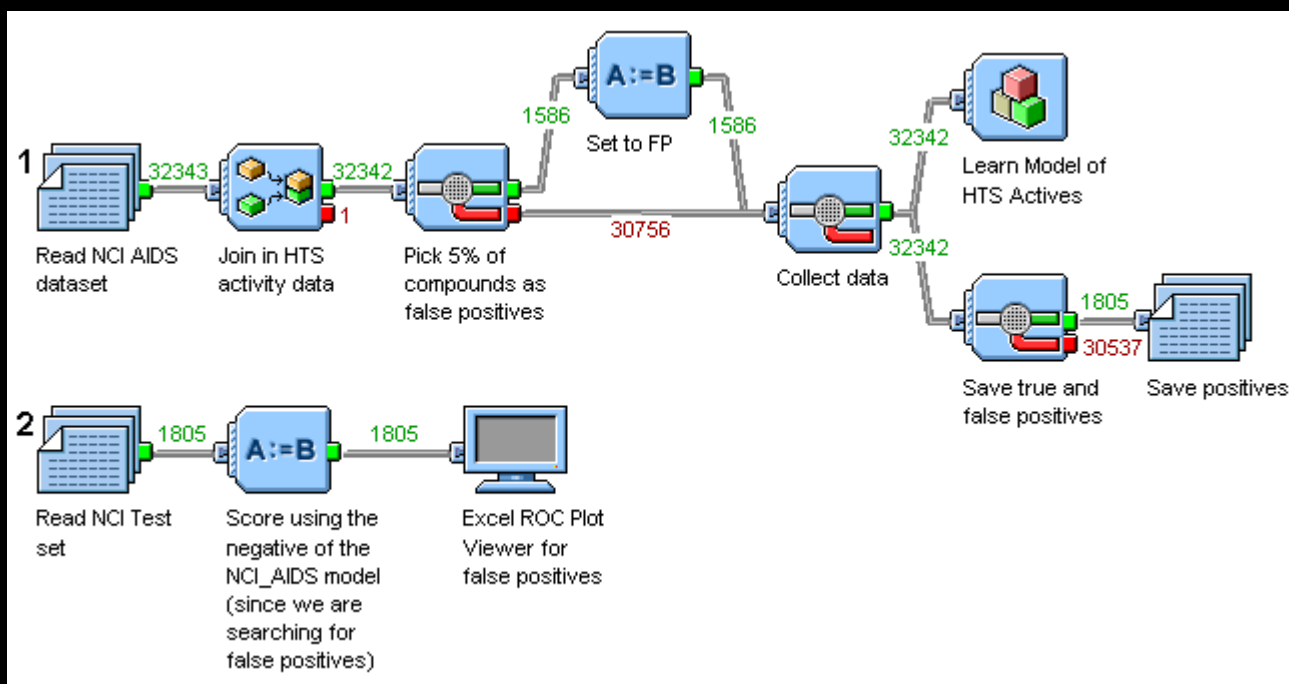
Results

- Results:
 - Weak actives aided in discovery of highly-active compounds
- (Side note: the enrichment curve for the weak actives left in the test set was *worse* than that of the high actives!)

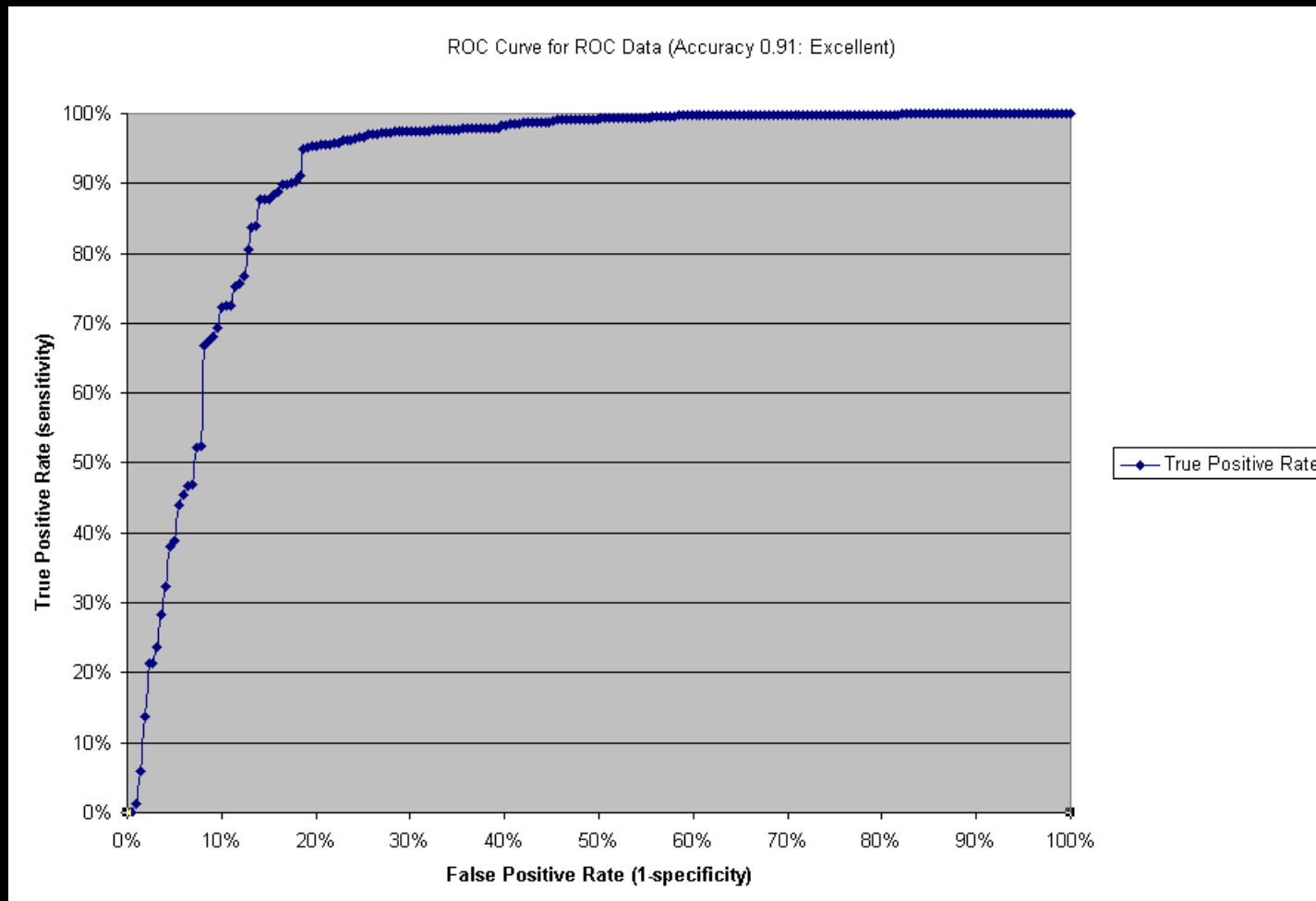


Experiment 5: Search for false positives

- Confirmation of hits expensive
- Retest often the slowest part of screening
- Train on data containing 12% true actives, 88% false positives

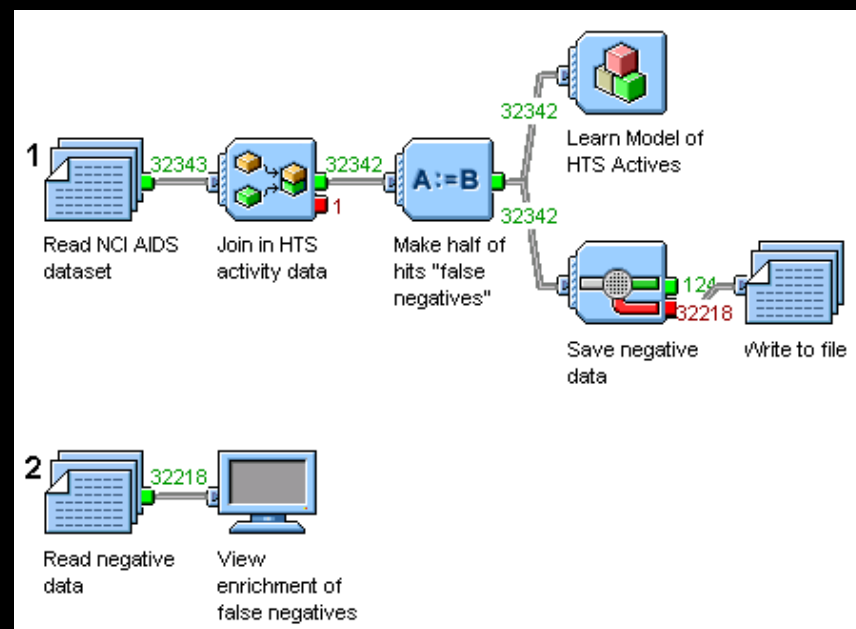


Results



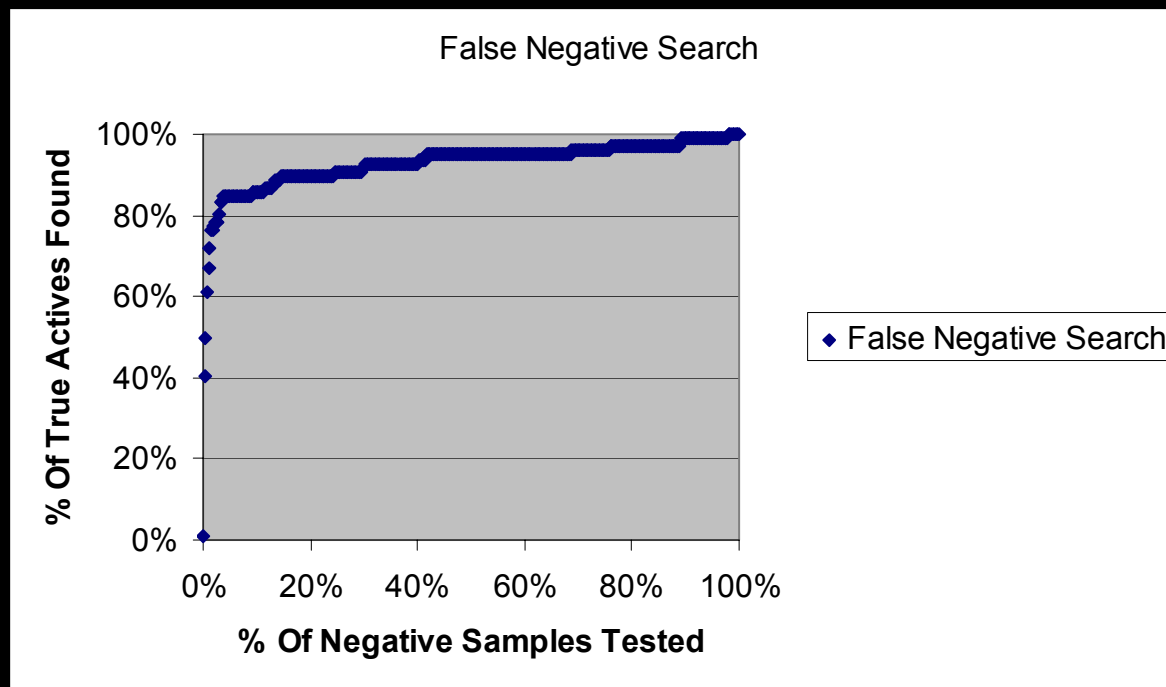
Experiment 6: Search for false negatives

- False negatives problematic
 - Even more costly to retest negatives
 - Can disrupt SAR studies
- Experiment:
 - Take half of 230 hits and mark them as inactive
 - Build model with data set
 - Sort negatives for retest



Results

- 85% found in top 5% of negatives



Conclusion

- Bayesian method resilient to many common HTS stresses
- Pipeline Pilot good testbed for computational experiments