



# Modeling HTS Data for Activity Prediction and Iterative Screening

Robert Brown, PhD and David Rogers, PhD  
SciTegic Inc

LabAutomation 2005, San Jose, CA

# Tasks in modeling HTS data

- Data collection
  - Assemble test and training set of molecules with response variable(s)
- Model building
  - Identify relationships between structure and response
    - Encode the structure as descriptors
    - Apply a statistical method to discover the relationship between the descriptors and the response
- Model application
  - Predict the activity of molecules proposed for synthesis or acquisition using the model(s)

# Challenges in modeling HTS data

- Method must be robust to
  - Large volumes of data
  - Skewed occurrence of classes with low occurrence of the interesting class (i.e. the hits!)
  - Noise – both false positives and false negatives
  - Multiple modes of activity

# Outline

- **Methods**
  - Extended connectivity fingerprints
  - Bayesian learning
- **Case study**
  - Data mining the NCI AIDS data set
  - Simulating screening prioritization
- **Ongoing work**
  - Modeling selectivity – capturing multiple responses in the same model
    - e.g activities and side effects

# Extended Connectivity Fingerprints (ECFP)

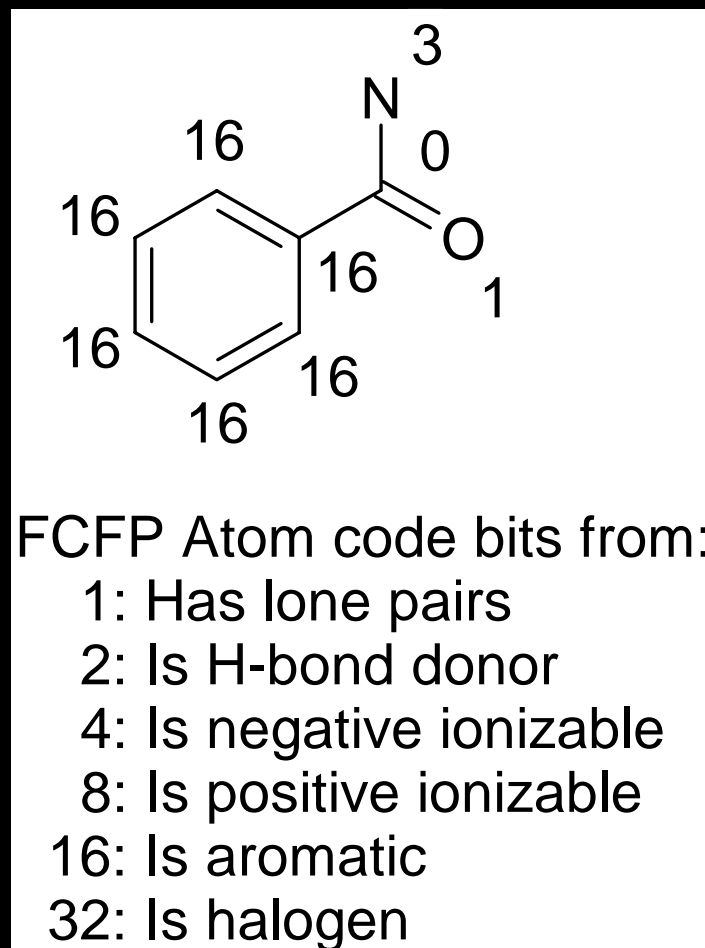
- A new descriptor for molecular characterization
- Goals of the fingerprint
  - Be **comprehensive** – encode “all” features within a structure
    - do not rely on a pre-defined dictionary of features
    - encode tertiary/quaternary information (c.f. path fingerprints)
    - encode substitution patterns to the fragment
  - Create an **interpretable** model
    - Each bit in the fingerprint should represent a single decodable feature
  - Be **fast** to calculate
    - Model building and especially virtual screening should be fast processes

# The FP Generation Process

- Process based on the Morgan algorithm
  - One of the first methods developed for computational chemistry
- Each atom is given an initial atom code
  - ECFP: Specific atom typing
  - **FCFP: Abstract functional role of atom**
- A number of iterations are performed
  - Each atom collects information from its neighbors
    - N iterations define structures 2N bonds wide
  - Resulting feature is mapped into a  $2^{32}$  address space

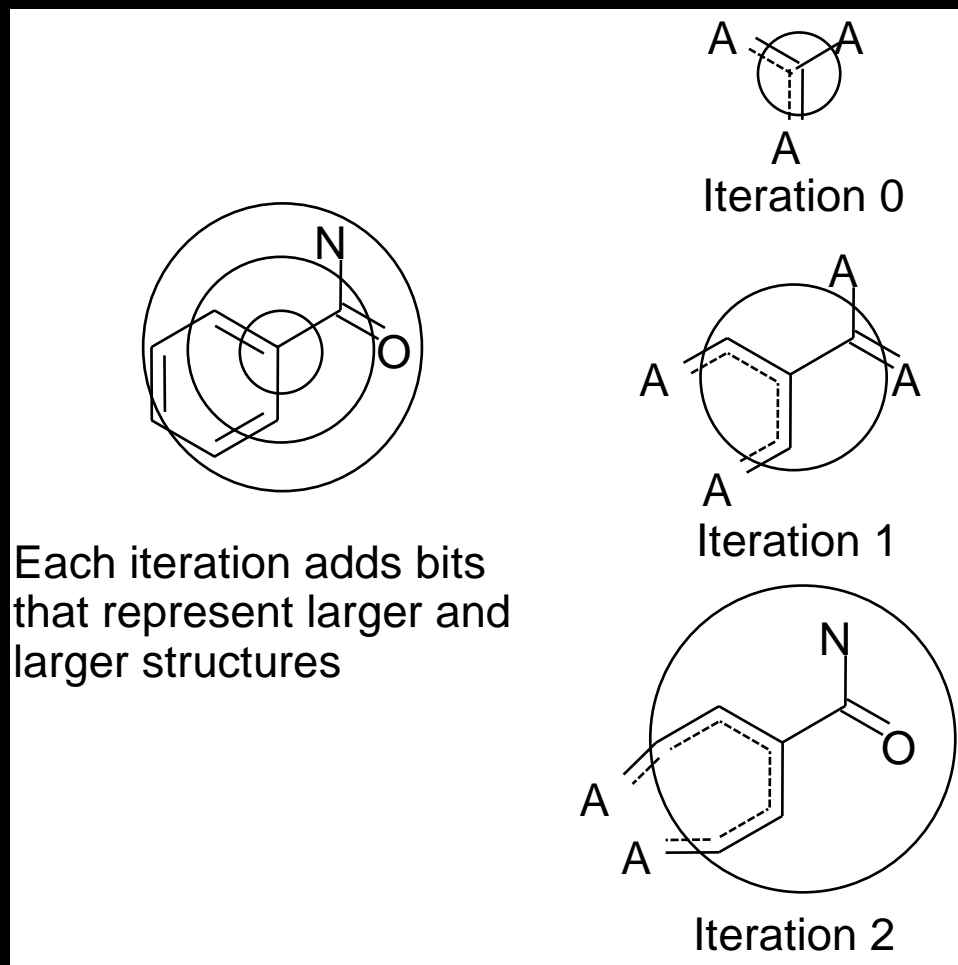
# FCFP: Functional-Class Fingerprints

- Use the role of an atom in the initial atom code rather than the atom type
  - Halogens give the same code
  - Hydrogen bond donors equivalent
  - Hydrogen bond acceptors equivalent



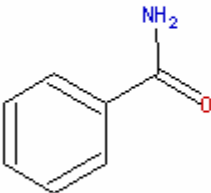
# Extending the initial atom codes

- Record (bond-type, atom-type) codes for each neighbour
- Sort to avoid order dependency
- Apply hashing function to map to a single number in the  $2^{32}$  address space (~4 billion bits)
- Chance of collisions is *extremely* low



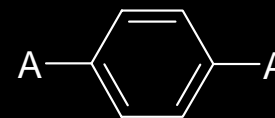
# FCFP: Generating the Fingerprint

- Information gain peaks after a few iterations

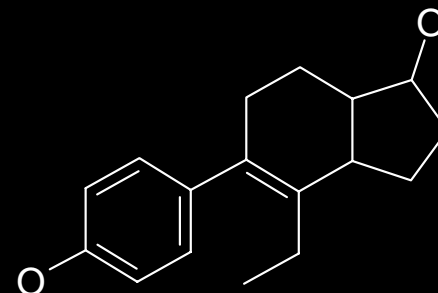
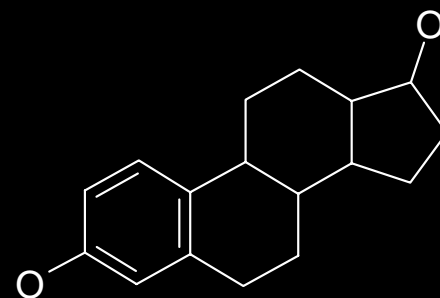
Molecule	FCFP_0	FCFP_2	FCFP_4	FCFP_6	FCFP_8
	16	16	16	16	16
	0	0	0	0	0
	1	1	1	1	1
	3	3	3	3	3
	1618154665	1618154665	1618154665	1618154665	1618154665
	203677720	203677720	203677720	203677720	203677720
	-1549103449	-1549103449	-1549103449	-1549103449	-1549103449
	1872154524	1872154524	1872154524	1872154524	1872154524
	1070061035	1070061035	1070061035	1070061035	1070061035
	991735244	991735244	991735244	991735244	991735244
	-453677277	-453677277	-453677277	-453677277	-453677277
	-581879738	-581879738	-581879738	-581879738	-581879738
	-1094243697	-1094243697	-1094243697	-1094243697	-1094243697
	-1698724694	-1698724694	-1698724694	-1698724694	-1698724694
	-2093839777	-2093839777	-2093839777	-2093839777	-2093839777
	380513738	380513738	380513738	380513738	380513738

# ECFPs and FCFPs

- New class of fingerprints for molecular characterization
  - Each bit represents the presence of a structural (not substructural) feature
  - Multiple levels of abstraction contained in single FP
- Large but sparse
  - Typical molecule generates 100s - 1000s of bits
  - Typical library generates 100K - 10M different bits.
- Fast
  - Generated at 10,000 mols/sec (2GHz PC)
  - Tanimoto pairwise similarities at ~500K comparisons/sec



Feature



# Outline

- Challenges in the analysis of HTS data
- Methods
  - Extended connectivity fingerprints
  - **Bayesian learning**
- Case study
  - Data mining the NCI AIDS data set
  - Simulating screening prioritization

# 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)
- Ideal for vHTS applications
  - Efficient:
    - Fast & 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

# An example model

	A	B	C	D	E	F
1	Equation "NCI_AIDS"					
2	Features from: ("FCFP_6" LongFingerprintType)					
3	Features from: ("ALogP" DoubleType)					
4	Features from: ("Molecular_Weight" DoubleType)					
5	Features from: ("Num_H_Donors" LongType)					
6	Features from: ("Num_H_Acceptors" LongType)					
7	Features from: ("Num_RotatableBonds" LongType)					
8						
9	Feature Statistics:					
10						
11	Property "FCFP_6":					
12	Total # of features in all samples: 710168 in subset: 6250					
13						
14	POSITIVE BINS					
15	Bin ID	G1	G2	G3	G4	G5
16	Bin Value	156149520	-353280459	-1838036449	-1.216E+09	-1804091252
17	Feature Co	31	32	32	39	113
18	Subset Co	24	24	24	24	36
19	Normalized	2.977639	2.970748	2.970748	2.9238	2.920533
20						

NEGATIVE BINS						
Bin ID	B1	B2	B3	B4	B5	
63	Bin Value	-387072142	-885520711	-1078052987	581019816	135188430
66	Feature Co	1079	913	858	836	620
67	Subset Co	0	0	0	0	0
68	Normalized	-2.350994	-2.201114	-2.146052	-2.123149	-1.86508
69						

395						
396	Property "Molecular_Weight":					
397	Total # of features in all samples: 16010 in subset: 114					
398						
399	POSITIVE BINS					
400	Bin ID	G1	G2	G3	G4	
401	Bin Minimum	1145.75371	690.89359	1032.038683	1373.18378	
402	Bin Maximum	1259.46875	804.608621	1145.753714	1486.89881	
403	Feature Co	23	266	42	13	
404	Subset Co	4	8	3	2	
405	Normalized	1.457771	1.134562	1.124651	1.010082	
406						

**Class: Good features from FCFP\_6**

**Class: Bad features from FCFP\_6**

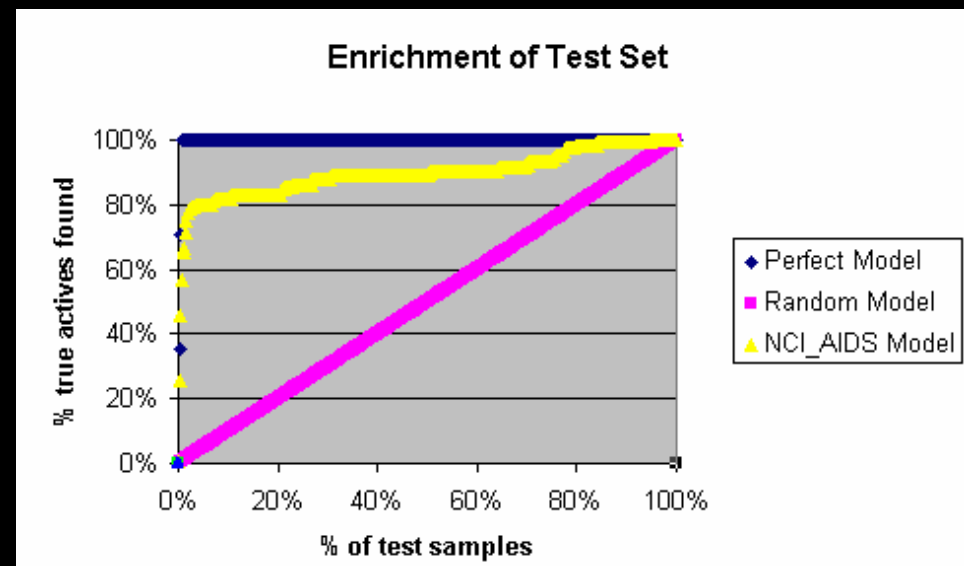
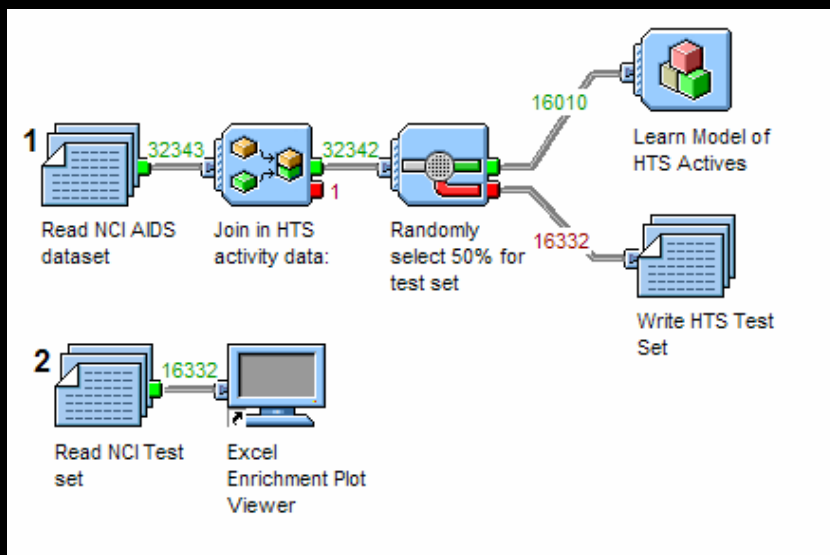
The figure displays two sets of chemical structures. The top set, titled "Class: Good features from FCFP\_6", shows eight structures (G1-G8) with their respective bin IDs and performance metrics on good samples. The bottom set, titled "Class: Bad features from FCFP\_6", shows four structures (B1-B4) with their respective bin IDs and performance metrics on good samples. The structures are: G1: 156149520 (24 out of 31 good); G2: -353280459 (24 out of 32 good); G3: -1838036449 (24 out of 32 good); G4: -1216241273 (24 out of 39 good); G9: -251375316 (21 out of 26 good); G10: 1445292113 (24 out of 46 good); G11: -1281108834 (21 out of 32 good); G12: 160164540 (24 out of 62 good); B1: -387072142 (0 out of 1079 good); B2: -885520711 (0 out of 913 good); B3: -1078052987 (0 out of 858 good); B4: 581019816 (0 out of 836 good).

## Case Study: NCI AIDS data

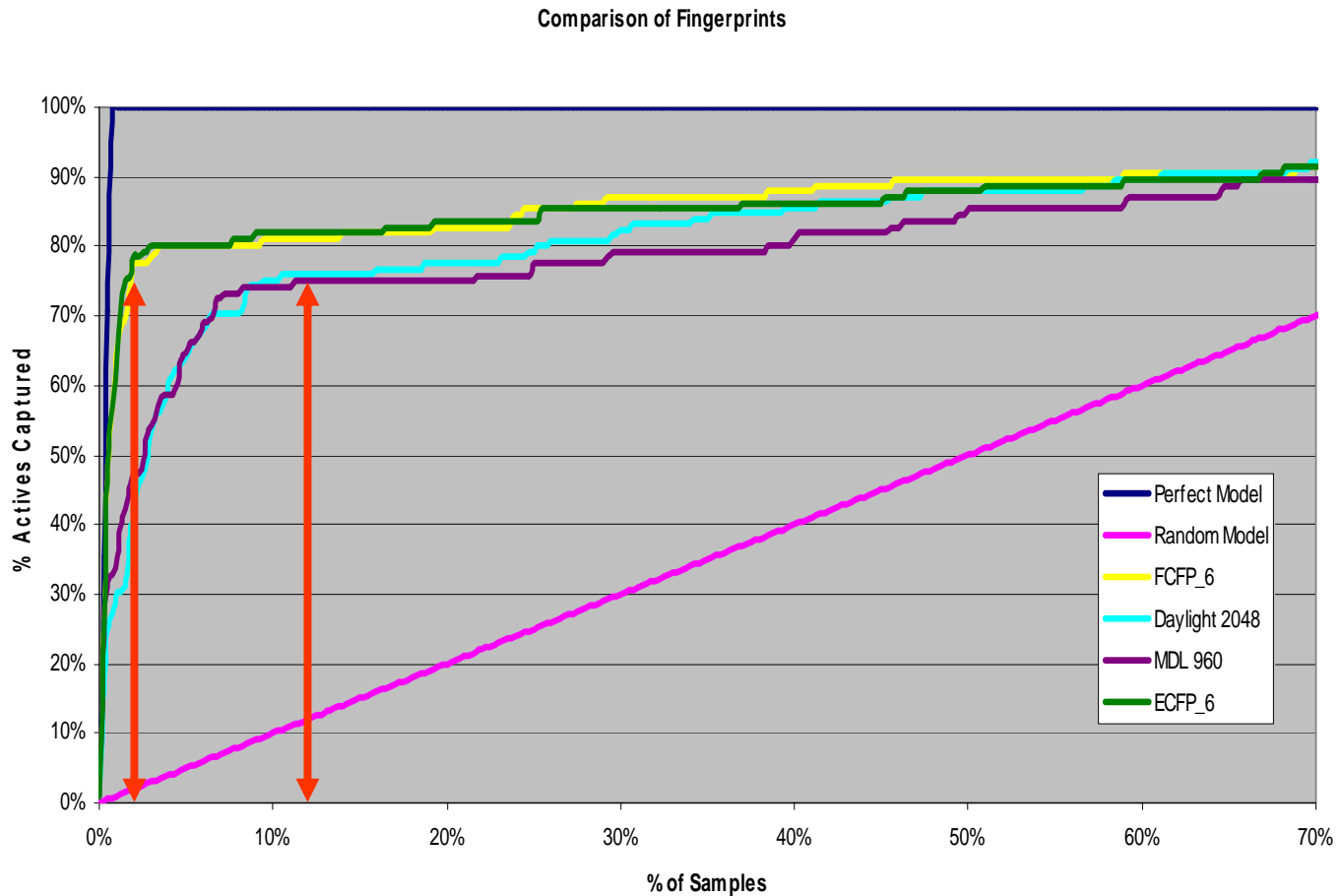
- ~32,000 compounds selected for HTS
- Whole-cell assay
- Found 230 confirmed hits (“CA”)
- Represent 7 “activity classes” (modes of activity)

# Results of Bayesian modeling

- Data split 50/50
  - Trained on 16,000 samples w/ 115 hits
  - FCFP\_6, AlogP, MW, #HBA, #HBD, #Rot Bonds
- Results:
  - Would have discovered 80% of actives screening ~600 cmpds
  - Model learned multiple modes of activity

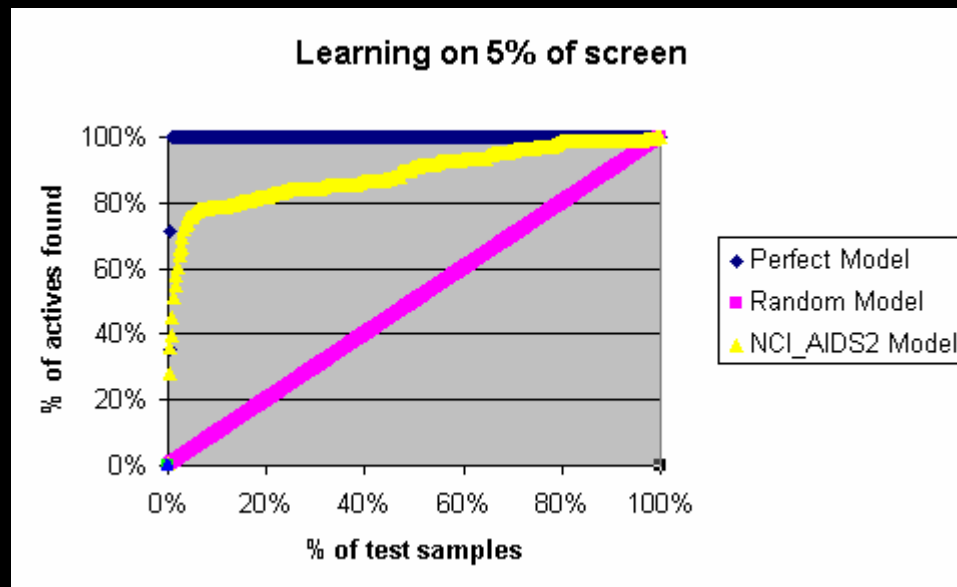


# Comparison of Fingerprints



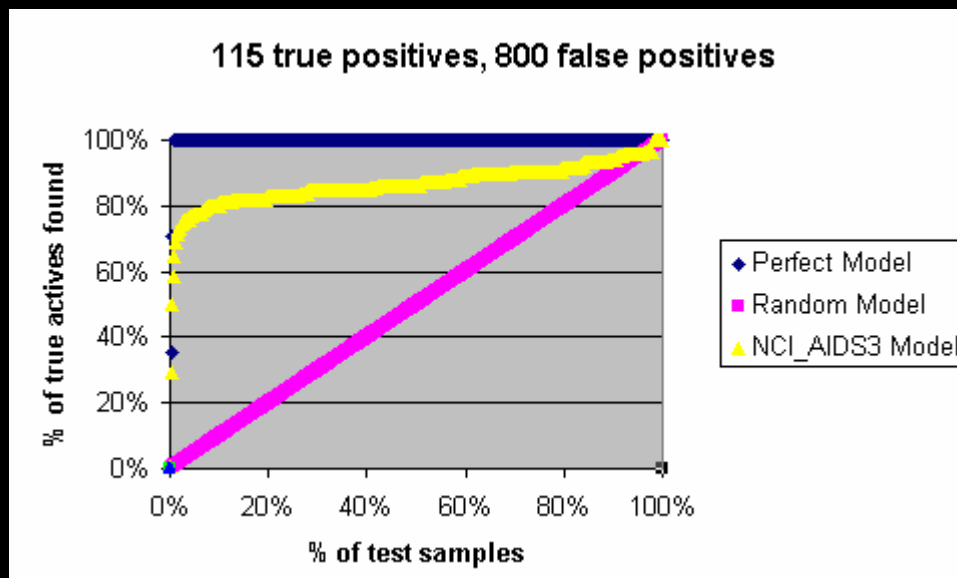
# Robust to small numbers of hits

- Data split 5/95
  - Trained on ~1,600 samples, 14 hits
- Results:
  - Would have discovered 80% of actives screening ~3,000 cmpds



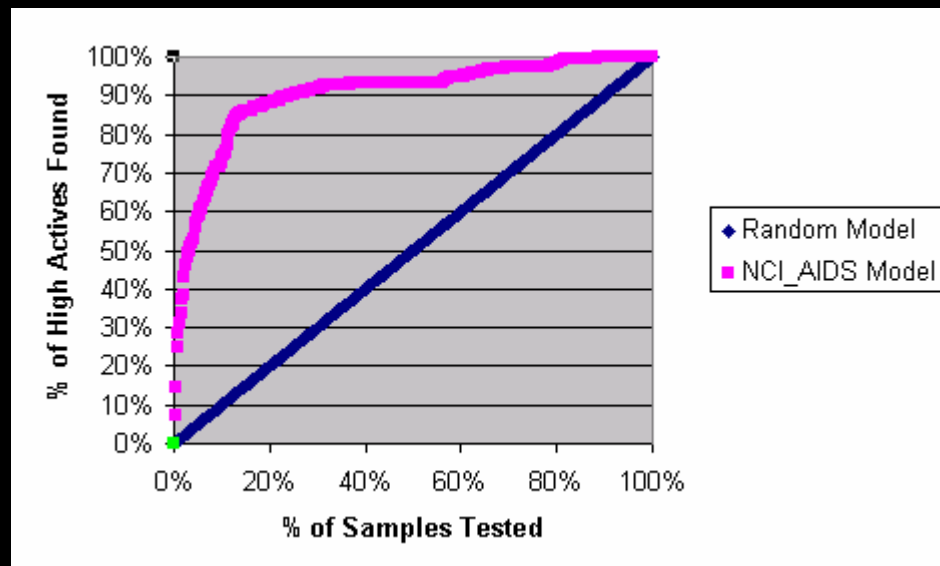
# Robust to noise in hits

- Data split 50/50
  - 5% of negatives in training set reassigned as *false positives*
  - Data contained 115 true actives and ~800 false actives
- Results:
  - Would have discovered 80% of actives screening ~1,500 cmpds



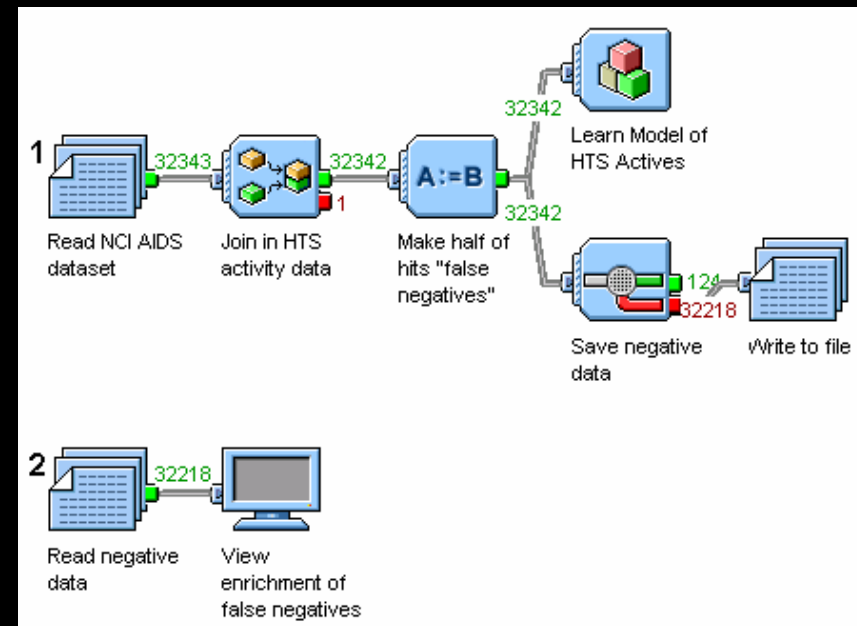
# Robust to weak actives for training

- Data split 50/50
  - All confirmed actives (CA) removed to test set
  - Trained on 130 confirmed *moderately active* (CM) compounds
- Results:
  - Weak actives aided in discovery of highly-active compounds



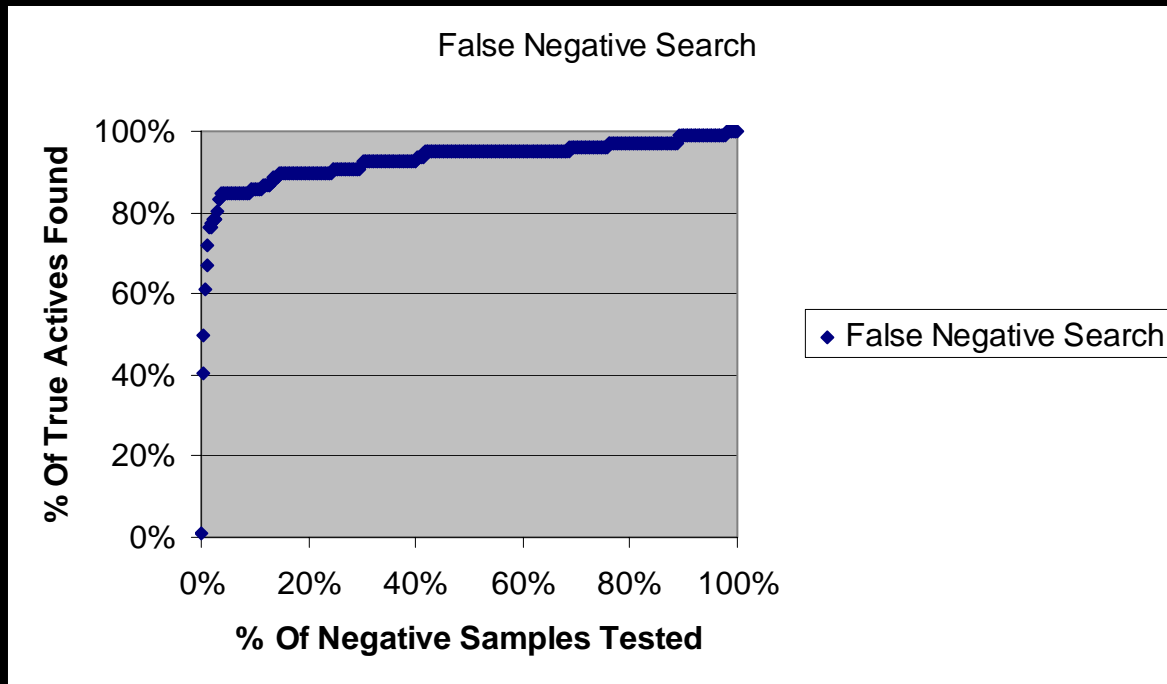
# Search for false negatives

- False negatives problematic
  - 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



# Search for false negatives

- 85% found in top 5% of negatives



# Outline

- Challenges in the analysis of HTS data
- Methods
  - Extended connectivity fingerprints
  - Bayesian learning
- Case study
  - Data mining the NCI AIDS data set
  - **Simulating screening prioritization**

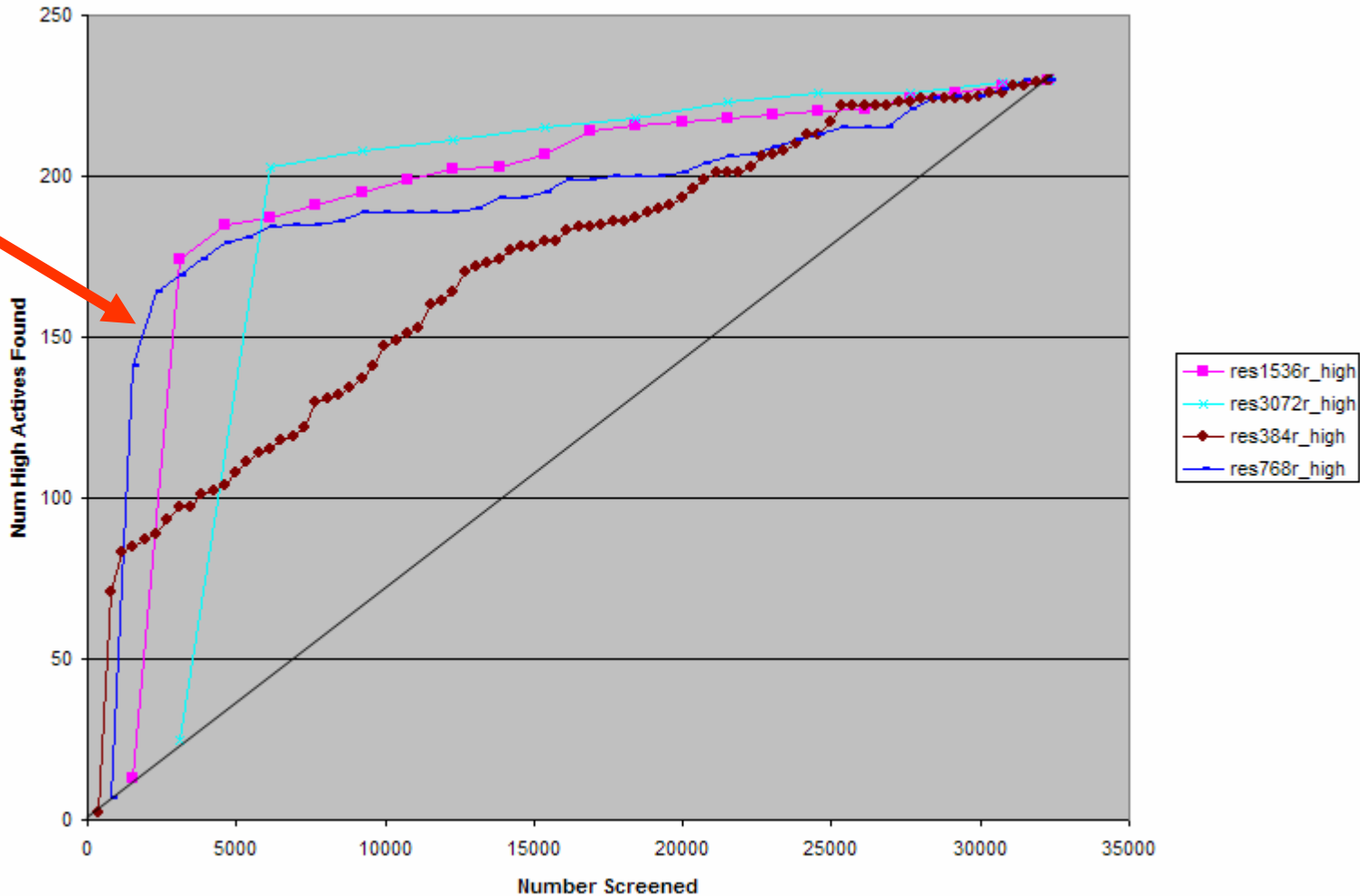
# Screening Prioritization

- HTS Screening strategies
  - Screen the entire compound collection
  - Iterative screening
    - Screen the entire collection in ordered subsets
    - Screen the collection in ordered subsets and stop when returns are diminishing Iterative screening
  - Screen a subset
    - Random / Ordered
    - Build a model of the screening results
    - Prioritize the remaining compounds and select the next subset to screen
    - Update the model and select the next subset
    - Repeat until
      - No more compounds
      - Hit rate falls below a set level

## Example

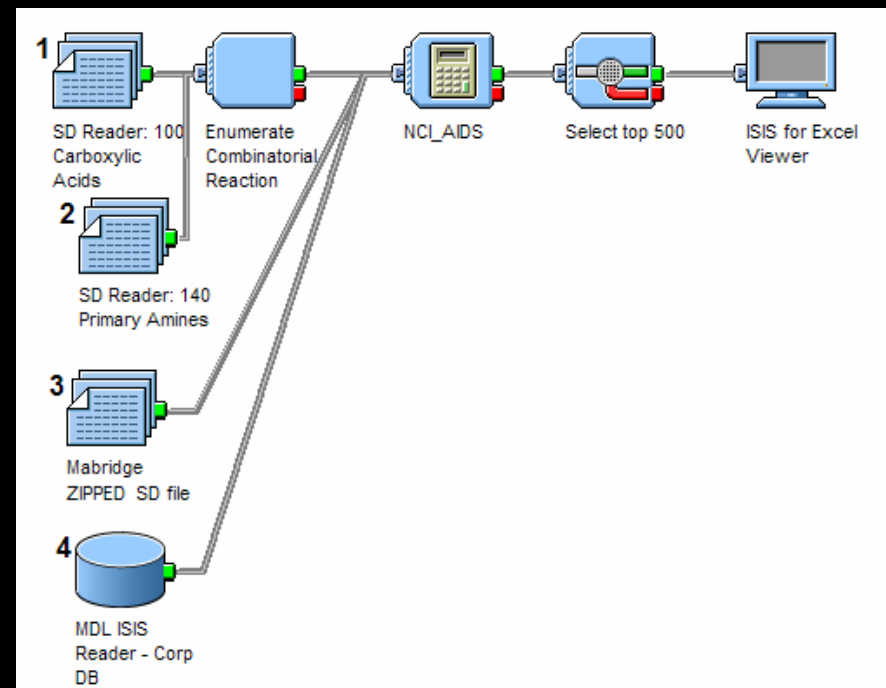
- Using the same NCI AIDS data set
  - Select a subset at random (384, 768, 1536, 3072)
  - “Screen” (i.e look up # actives)
  - Build a Bayesian model
  - Score the remaining compounds
  - Sort by score
  - Select the next subset of the same size and “screen”
  - Repeat until all molecules are “screened”

# Actives found per iteration

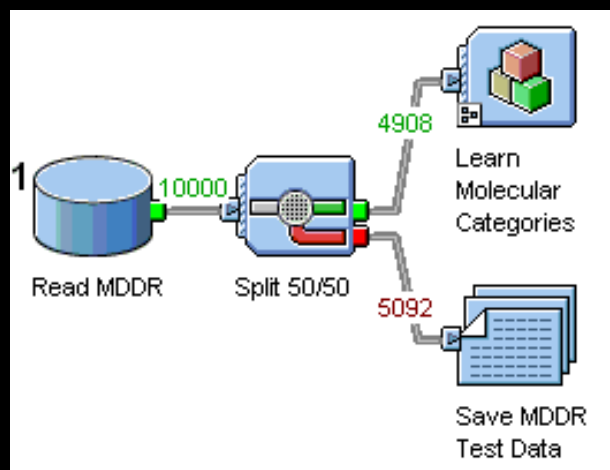


# Using the models

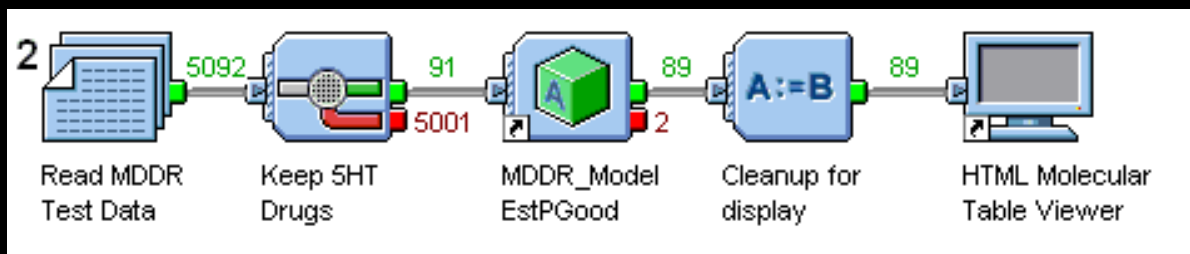
- Models can be used as virtual screens to filter
  - Virtual combichem libraries
  - Vendor files e.g Maybridge
  - Vendor databases e.g. ACD
  - Corporate databases



# Modeling Selectivity

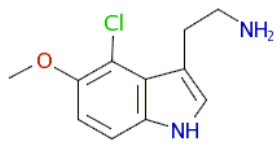
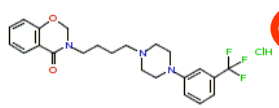
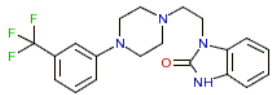


Modified Bayesian Method



Known activity

Normalized Predicted activity (0-1)

 <p>5 HT1D Agonist Antimigraine Antihypertensive</p>		<p>MDDR_Model_5 HT1D Agonist_EstPGood MDDR_Model_Melatonin Antagonist_EstPGood MDDR_Model_Antimigraine_EstPGood MDDR_Model_Melatonin Agonist_EstPGood MDDR_Model_Antihypertensive_EstPGood</p>	<p>0.989 0.971 0.813 0.307 0.111</p>
 <p>5 HT1A Agonist Anxiolytic Antidepressant Antihypertensive</p>		<p>MDDR_Model_5 HT1A Agonist_EstPGood MDDR_Model_5 HT2A Antagonist_EstPGood MDDR_Model_Antipsychotic_EstPGood MDDR_Model_Anxiolytic_EstPGood MDDR_Model_Antidepressant_EstPGood MDDR_Model_Adrenergic (alpha) blocker_EstPGood MDDR_Model_Antiarrhythmic_EstPGood</p>	<p>0.975 0.523 0.445 0.426 0.397 0.234 0.117</p>
 <p>5 HT1A Agonist 5 HT2A Antagonist Antidepressant</p>		<p>MDDR_Model_5 HT1A Agonist_EstPGood MDDR_Model_5 HT2A Antagonist_EstPGood MDDR_Model_Anxiolytic_EstPGood MDDR_Model_Antidepressant_EstPGood MDDR_Model_Antipsychotic_EstPGood</p>	<p>0.534 0.398 0.32 0.293 0.167</p>

## Summary

- New fingerprint for molecular characterization
  - Fast, comprehensive and interpretable
- Bayesian learning
  - Successfully model HTS data
  - Robust to low hit rate and noise
  - Able to identify false negatives for retest
- Screening prioritization
  - Can identify most actives early in a screen