



Organizing and Mining HTS Data using Data Pipelining

Robert Brown, David Rogers and Andrei Caracoti,
SciTegic Inc

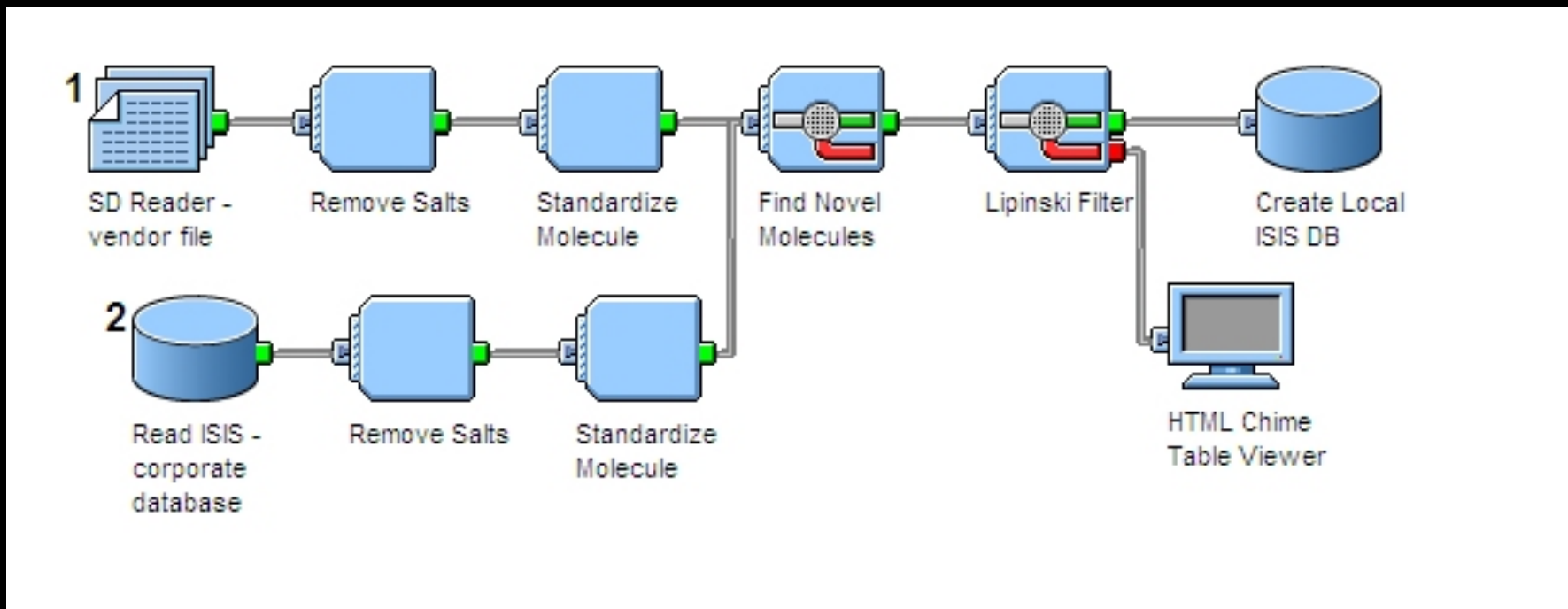
ALA LabFusion 2004, Boston

Outline

- Introduction to data pipelining
- Methods
 - Extended connectivity fingerprints
 - Bayesian learning
- Case study
 - Data mining the NCI AIDS data set
 - Simulating screening prioritization

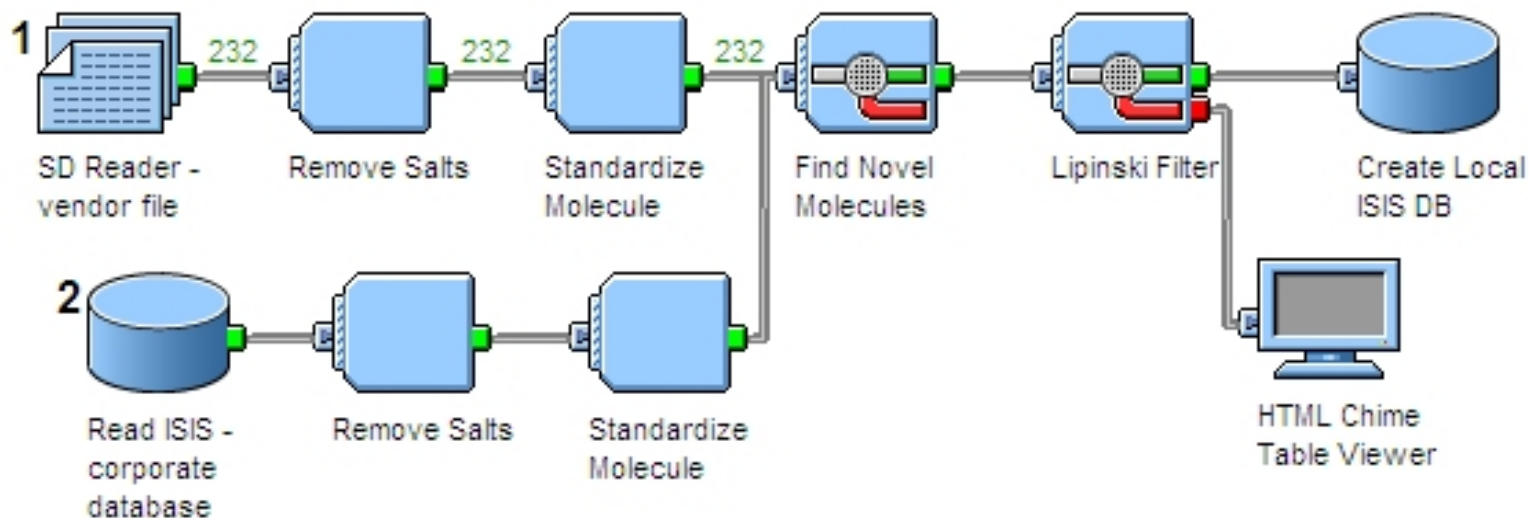
Data Pipelining

- A powerful new paradigm for data processing
- Pipelines guide the flow of data through a network of modular computational components



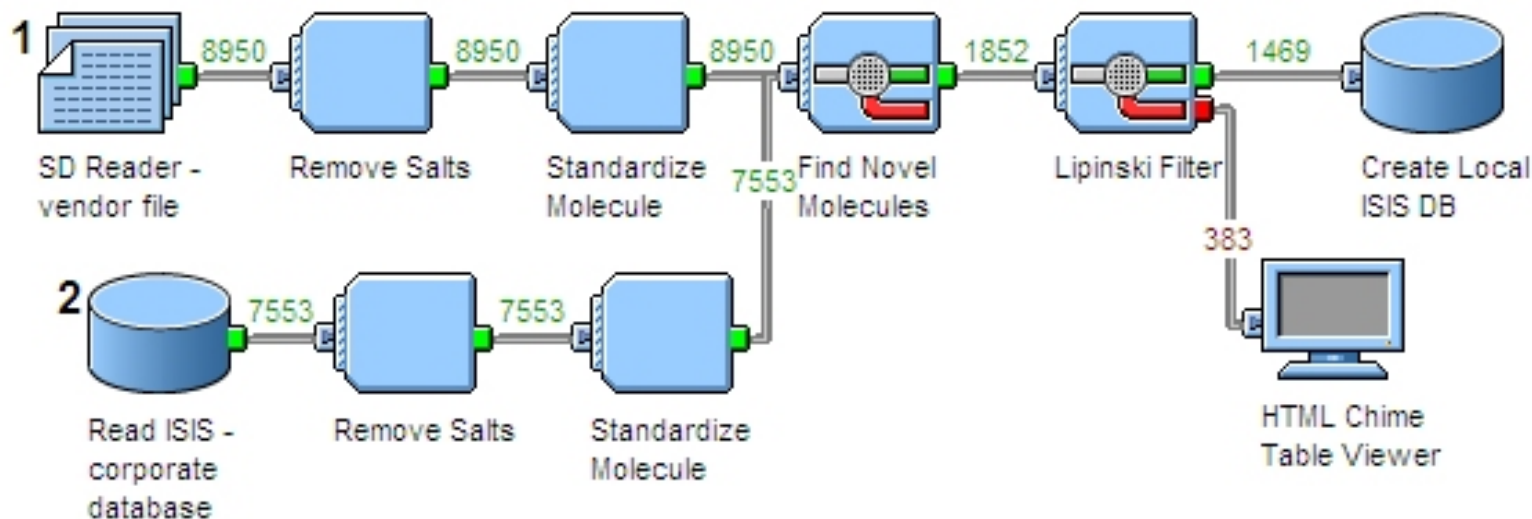
Data Pipelining

- A powerful new paradigm for data processing
- Pipelines guide the flow of data through a network of modular computational components



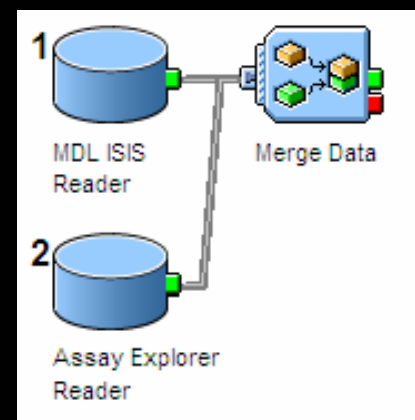
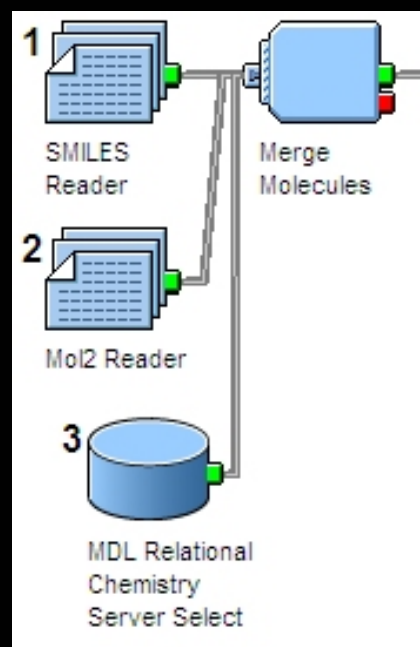
Data Pipelining

- A powerful new paradigm for data processing
- Pipelines guide the flow of data through a network of modular computational components



Data pipelining enables

- Processing of data from multiple disparate data sources
- Integration of disparate applications
- Rapid processing of large amounts of data
- Automated execution of routine processes
- Capture of best practice



Data pipelining enables

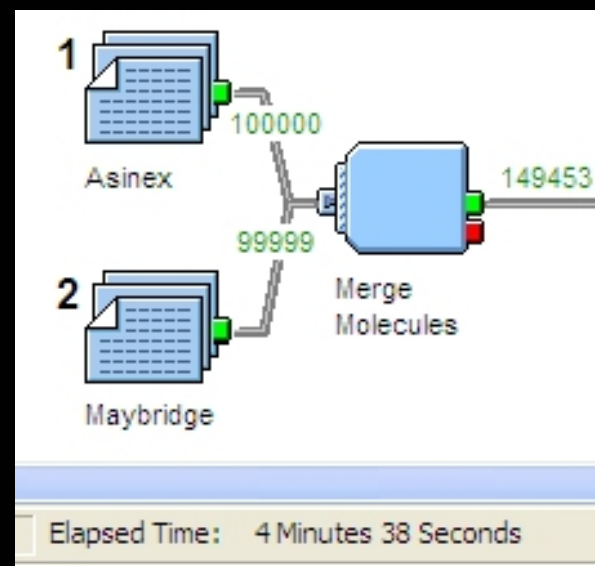
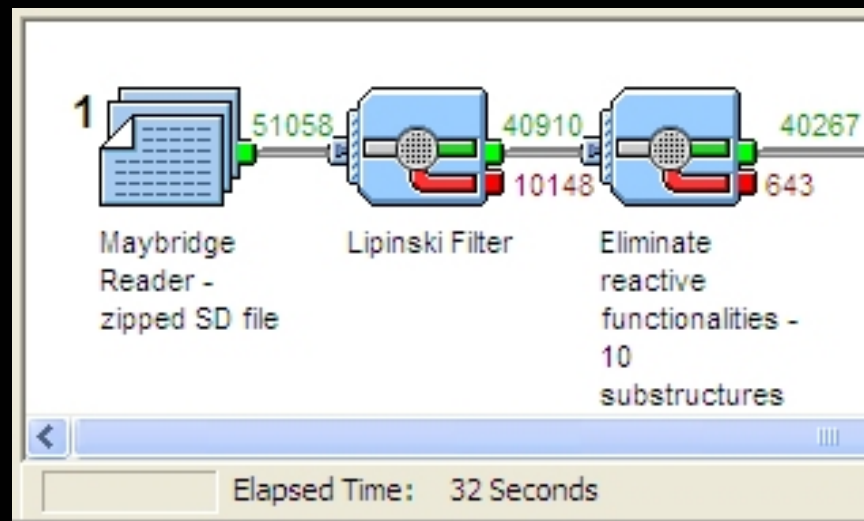
- Processing of data from multiple disparate data sources
- Integration of disparate applications
- Rapid processing of large amounts of data
- Automated execution of routine processes
- Capture of best practice



The Structure	CLogP	CLogP_MR	MDL_SS_Keys_166_N
	2.002	5.160	32
	1.372	3.591	28
	0.721	3.651	32

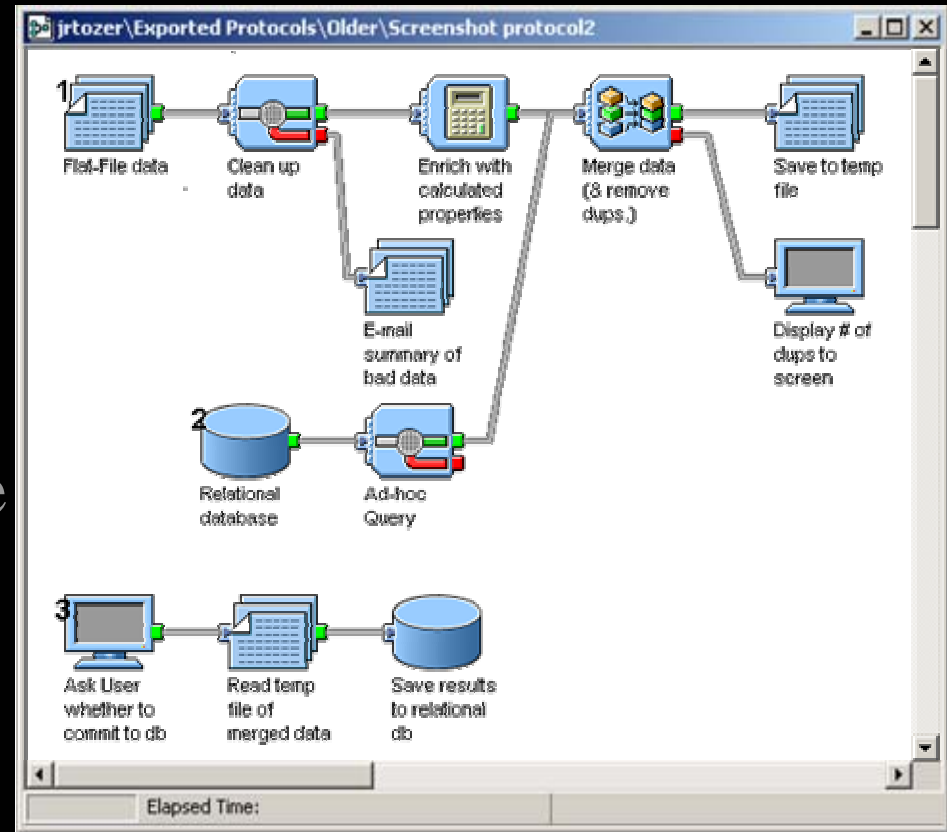
Data pipelining enables

- Processing of data from multiple disparate data sources
- Integration of disparate applications
- Rapid processing of large amounts of data
- Automated execution of routine processes
- Capture of best practice



Data pipelining enables

- Processing of data from multiple disparate data sources
- Integration of disparate applications
- Rapid processing of large amounts of data
- Automated execution of routine processes
- Capture of best practice



Outline

- Introduction to data pipelining
- Methods
 - Extended connectivity fingerprints
 - Bayesian learning
- Case study
 - Data mining the NCI AIDS data set
 - Simulating screening prioritization

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

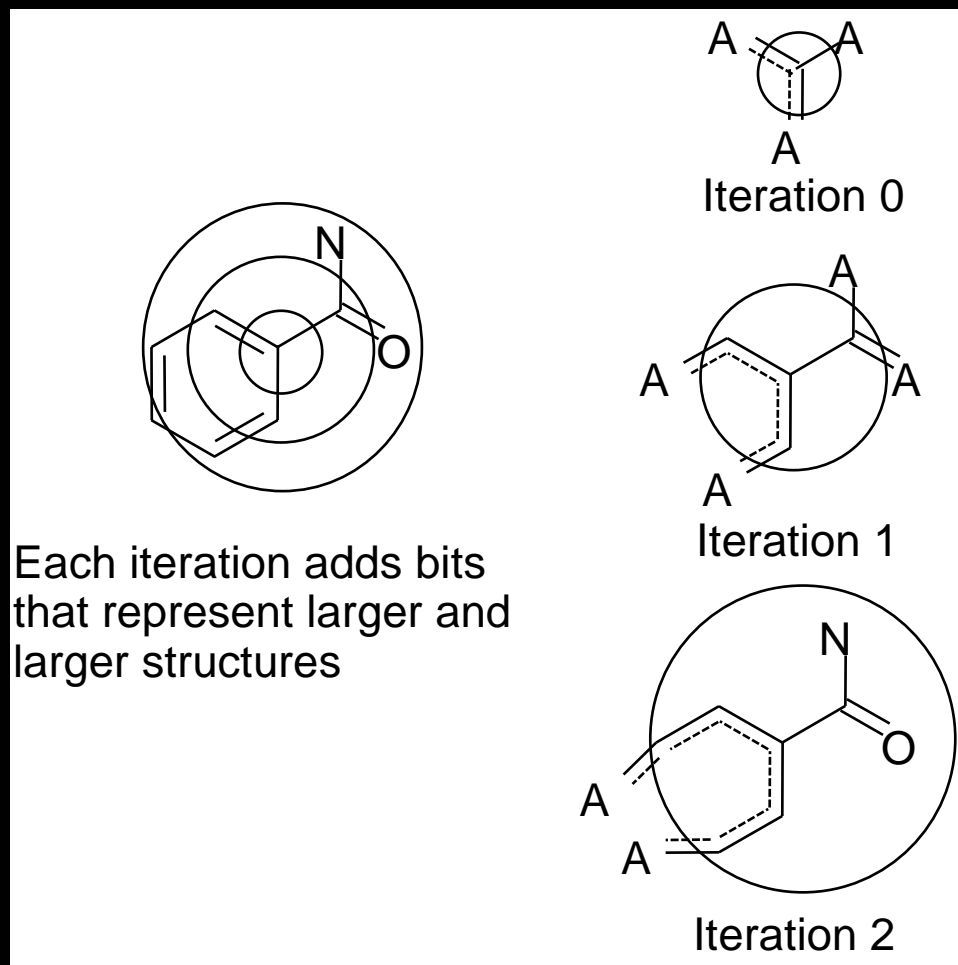
- Process based on the Morgan algorithm
 - One of the first methods developed for computational chemistry
- Each atom is given an initial atom code
 - ECFP: Specific atom typing
 - FCFP: Abstract functional role of atom
- A number of iterations are performed
 - Each atom collects information from its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

Assignment of Initial Atom Codes

- ECFPs
 - Atom type
 - Atom charge
 - Atom mass
 - Valence
 - Number of bond to non-hydrogens
 - Number of bonds to hydrogens
- Variant is to use the 120 AlogP atom types

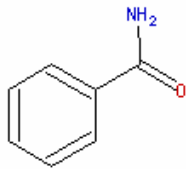
Extending the initial atom codes

- 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



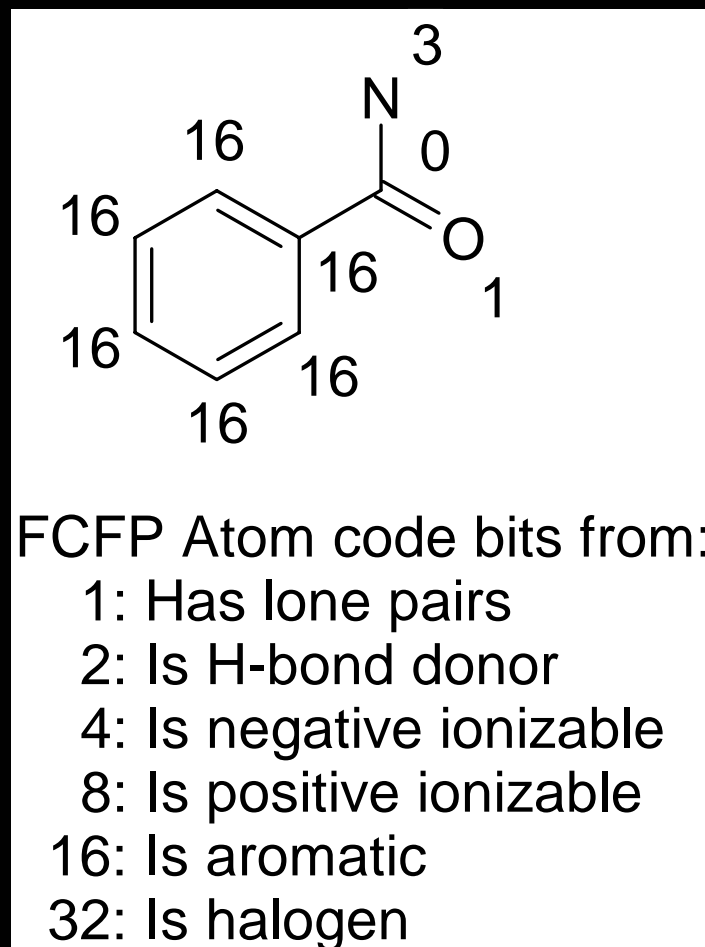
ECFP: Generating the Fingerprint

- Iteration is repeated desired number of times
- Codes from all iterations are collected
- Duplicate bits are removed
- Information gain diminishes after a few iterations

Molecule	ECFP_0	ECFP_2	ECFP_4	ECFP_6	ECFP_8
				-182236392	-182236392
				642810091	642810091
				-182236392	-182236392
				642810091	642810091
				-1100000244	-1100000244
				-1074141656	-1074141656
				-1100000244	-1100000244
				1572579716	1572579716
				-1074141656	-1074141656
				1572579716	1572579716
				1997021792	1997021792
				1996767644	1996767644
				1997021792	1997021792
				1996767644	1996767644
				-175146122	-175146122
				852414842	852414842
				2099970318	2099970318
			-932108170	-932108170	
			1564392544	1564392544	
			1571214559	1571214559	
			1451403962	1451403962	
			284029635	284029635	
			-948152242	-948152242	
			-1555299234	-1555299234	
			-281505363	-281505363	
			-344170121	-344170121	

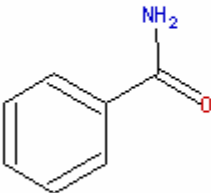
FCFP: Functional-Class Fingerprints

- 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



FCFP: Generating the Fingerprint

- Again, the information gained by reaching out further diminishes.

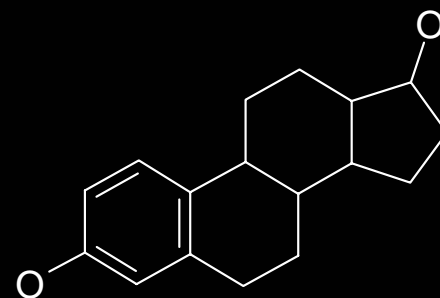
Molecule	FCFP_0	FCFP_2	FCFP_4	FCFP_6	FCFP_8
	16	16	16	16	16
	0	0	0	0	0
	1	1	1	1	1
	3	3	3	3	3
	1618154665	1618154665	1618154665	1618154665	1618154665
	203677720	203677720	203677720	203677720	203677720
	-1549103449	-1549103449	-1549103449	-1549103449	-1549103449
	1872154524	1872154524	1872154524	1872154524	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

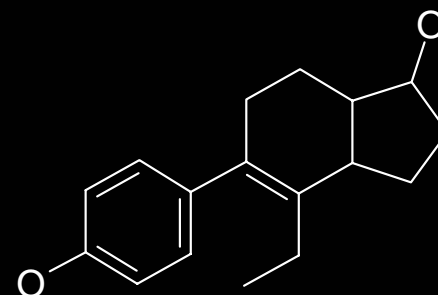
- New class of fingerprints for molecular characterization
 - Each bit represents the presence of a structural (not substructural) feature
 - Multiple levels of abstraction contained in single FP
- Large but sparse
 - Typical molecule generates 100s - 1000s of bits
 - Typical library generates 100K -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



Feature



X



✓

Outline

- Introduction to data pipelining
- 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

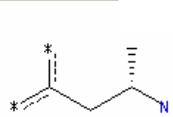
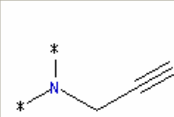
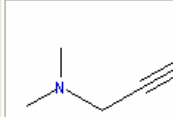
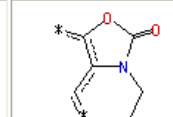
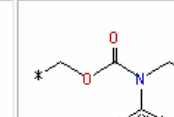
The Model

- Input is a training set with descriptors, a response variable and a test for good
- A *feature* is a binary attribute of a data record
 - For molecules: a *property range* or a *fingerprint bit*
- A count of each feature is kept:
 - Over all the samples
 - Over all samples that pass the test for good
- Normalized probability is calculated for **each feature**
 - $\log(\text{Laplacian corrected probability})$
- The normalized probabilities are summed over **all features** to give the *relative score*

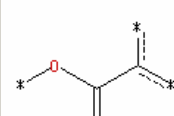
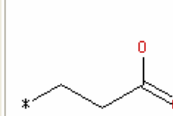
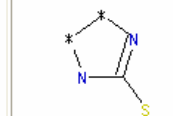
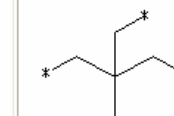
An example model

	A	B	C	D				
1	Equation "MAOInhibitorLike"							
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: 250988 in subset: 4004							
13								
14	POSITIVE BINS							
15	Bin ID	G1	G2	G3	G4	G5	G6	G7
16	Bin Value	-1290796621	-821137192	638618274	664924773	-4.1E+08	9.49E+08	1.97E+09
17	Feature Co	23	11	12	14	14	10	10
18	Subset Co	15	11	11	11	11	10	10
19	Normalized	2.46003	2.323228	2.309748	2.283321	2.283321	2.249881	2.249881

Class: Good features from FCFP_6

 G1: -1290796621 15 out of 23 good	 G3: 638618274 11 out of 12 good	 G4: 664924773 11 out of 14 good	 G5: -413588539 11 out of 14 good	 G8: -1304522383 10 out of 13 good
--	---	---	--	---

Class: Bad features from FCFP_6

 B1: 565968762 0 out of 769 good	 B2: -1549163031 0 out of 360 good	 B3: 394124770 0 out of 325 good	 B4: -1986098826 0 out of 324 good	 B5: -415245925 1 out of 699 good
--	---	---	---	--

50									
51	NEGATIVE BINS								
52	Bin ID	B1	B2	B3	B4	B5	B6	B7	B8
53	Bin Value	565968762	-1549163031	394124770	-1.986E+09	-4.2E+08	1.26E+09	65948508	5.6
54	Feature Co	769	360	325	324	699	294	279	
55	Subset Co	0	0	0	0	1	0	0	
56	Normalized	-2.585342	-1.908514	-1.82208	-1.819497	-1.80427	-1.73874	-1.69578	-1
57									

Normalized Probability

- Given a set of \underline{N} samples
- Given that some subset \underline{A} of them are good ('active')
 - Then we estimate for a new compound: $P(\text{good}) \sim A / N$
 - For a new feature to the model, this is our base estimate.
- Given a set of binary features F_i
 - For a given feature F :
 - It appears in N_F samples
 - It appears in A_F good samples
 - *Can we estimate* $P(\text{good} | F) \sim A_F / N_F$?
 - Different features are sampled different numbers of times
 - Error gets worse as $N_F \rightarrow$ small

Normalized Probability

- Solution: renormalize probabilities to baseline
 - Can be thought of as adding a single sample at baseline probability
- $P'(\text{good} | F) = (AF + P(\text{good})K) / (NF + K)$
 - $P'(\text{good} | F) \rightarrow P(\text{good})$ as $NF \rightarrow 0$
 - Assume: Most features have no relationship with activity
 - $P'(\text{good} | F) \rightarrow AF / NF$ as $NF \rightarrow \text{large}$
 - Assume: more instances of the observation the more likely it is not an artifact
- If $K = 1/P(\text{good})$ this is the Laplacian correction

Outline

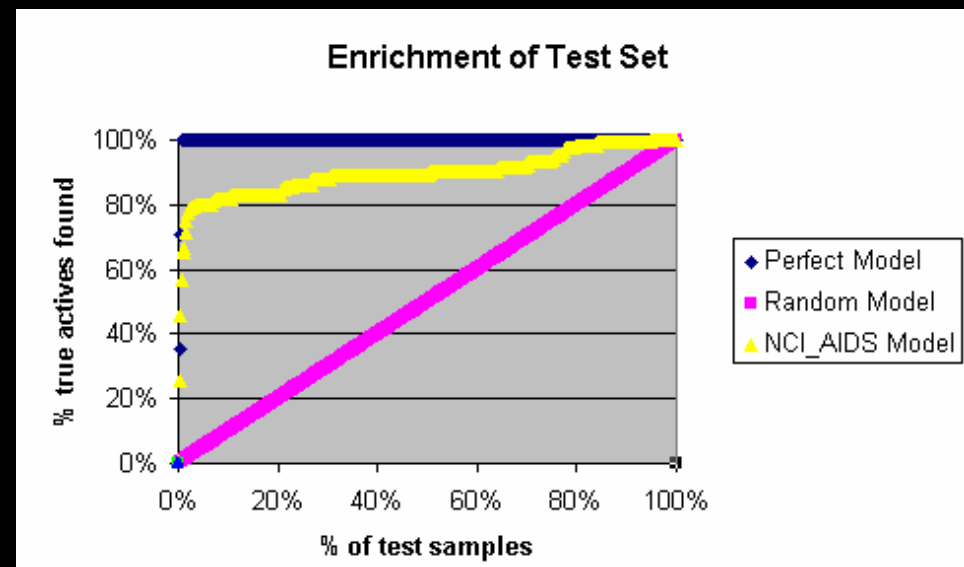
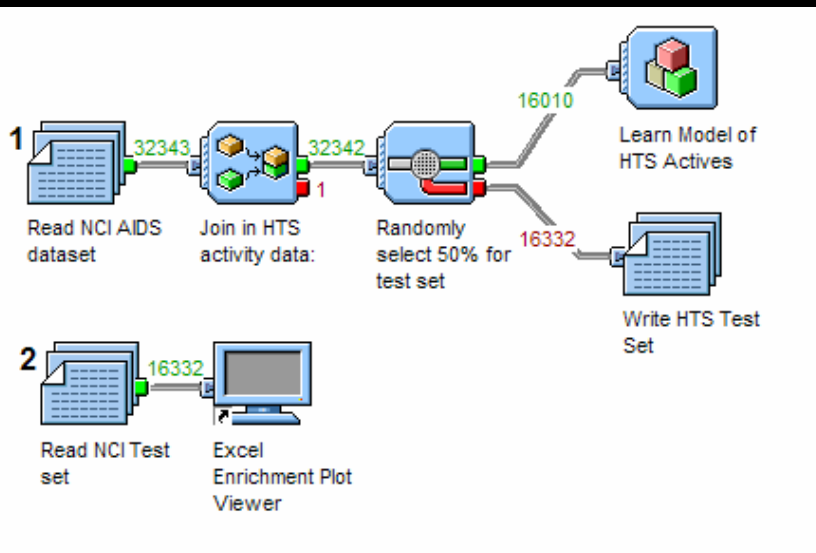
- Introduction to data pipelining
- Methods
 - Extended connectivity fingerprints
 - Bayesian learning
- Case study
 - **Data mining the NCI AIDS data set**
 - Simulating screening prioritization

Case Study: NCI AIDS data

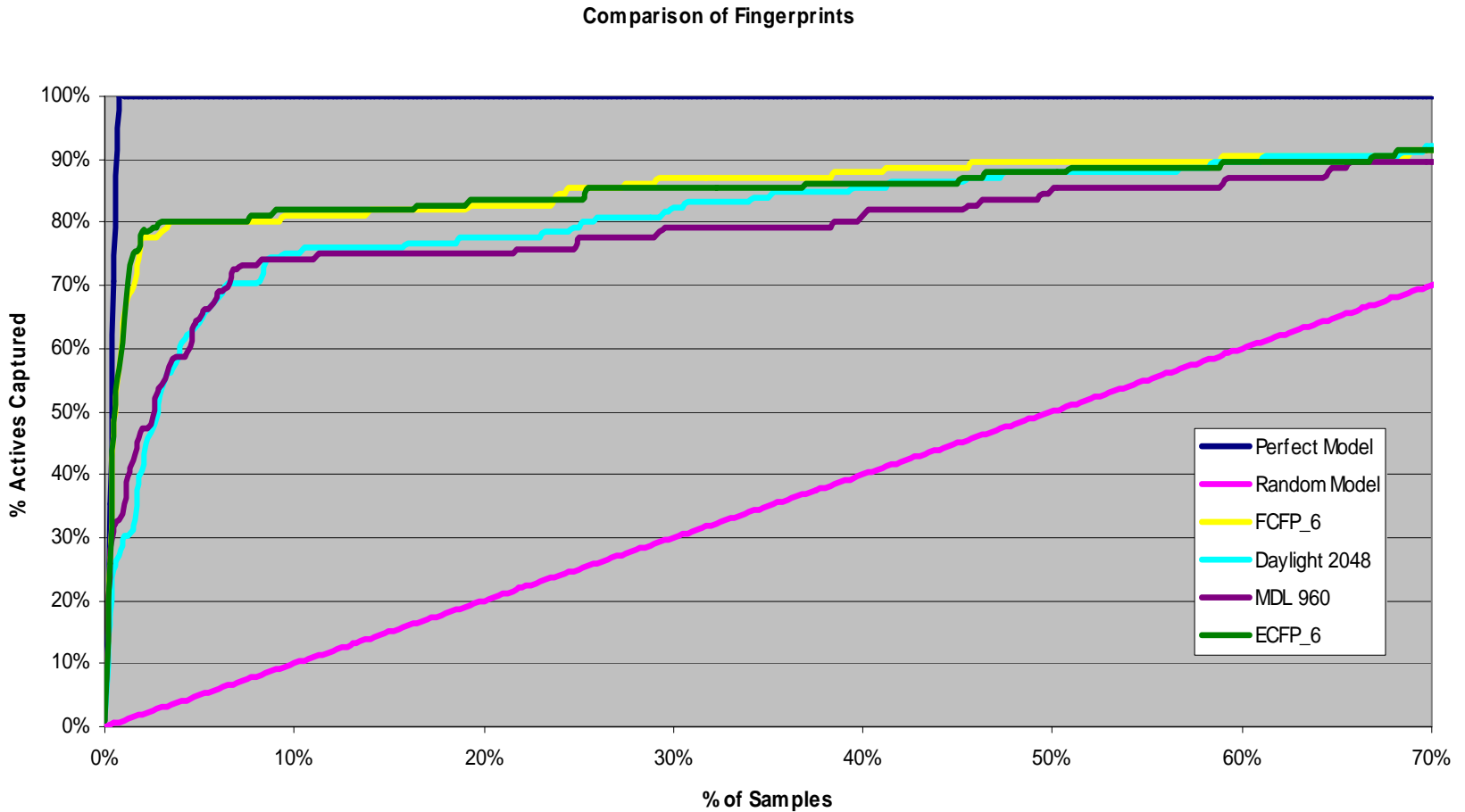
- ~32,000 compounds selected for HTS
- Whole-cell assay
- Found 230 confirmed hits (“CA”)
- Represent 7 “activity classes” (modes of activity)

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

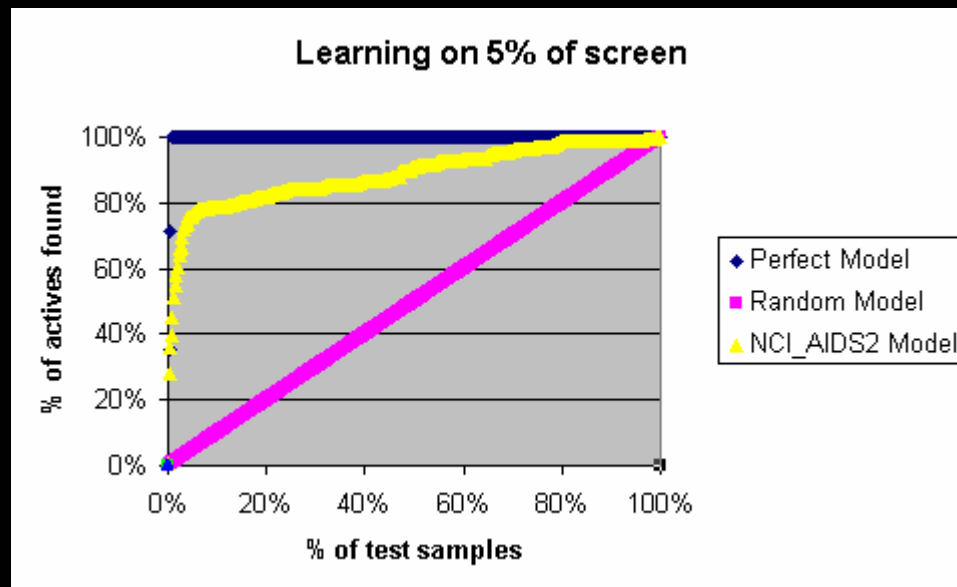


Comparison of Fingerprints



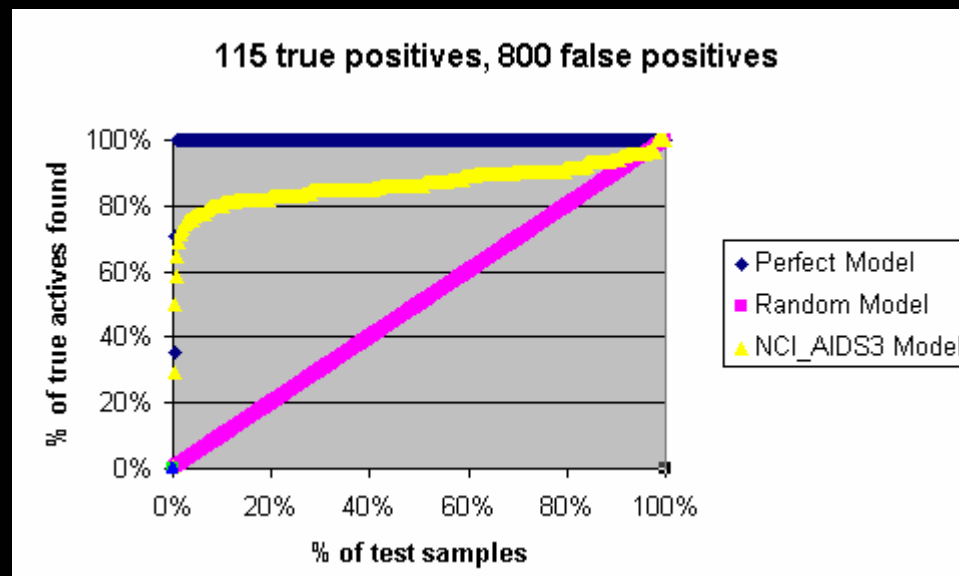
Robust to small numbers of hits

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



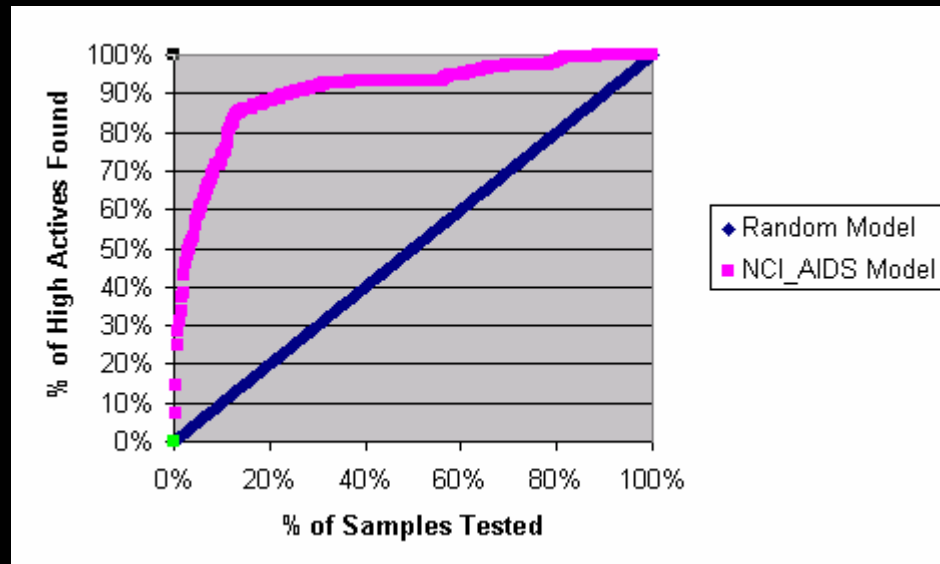
Robust to noise in hits

- 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



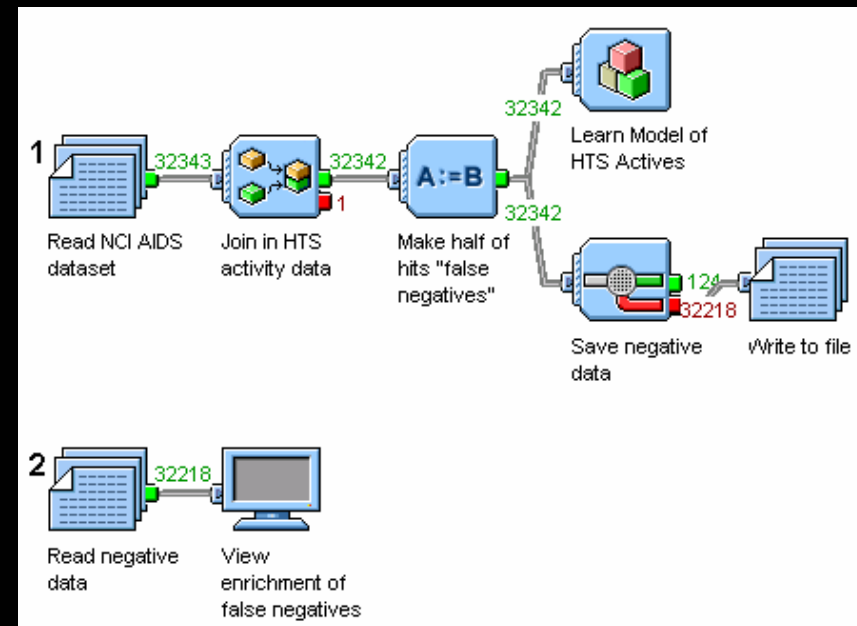
Robust to weak actives for training

- 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



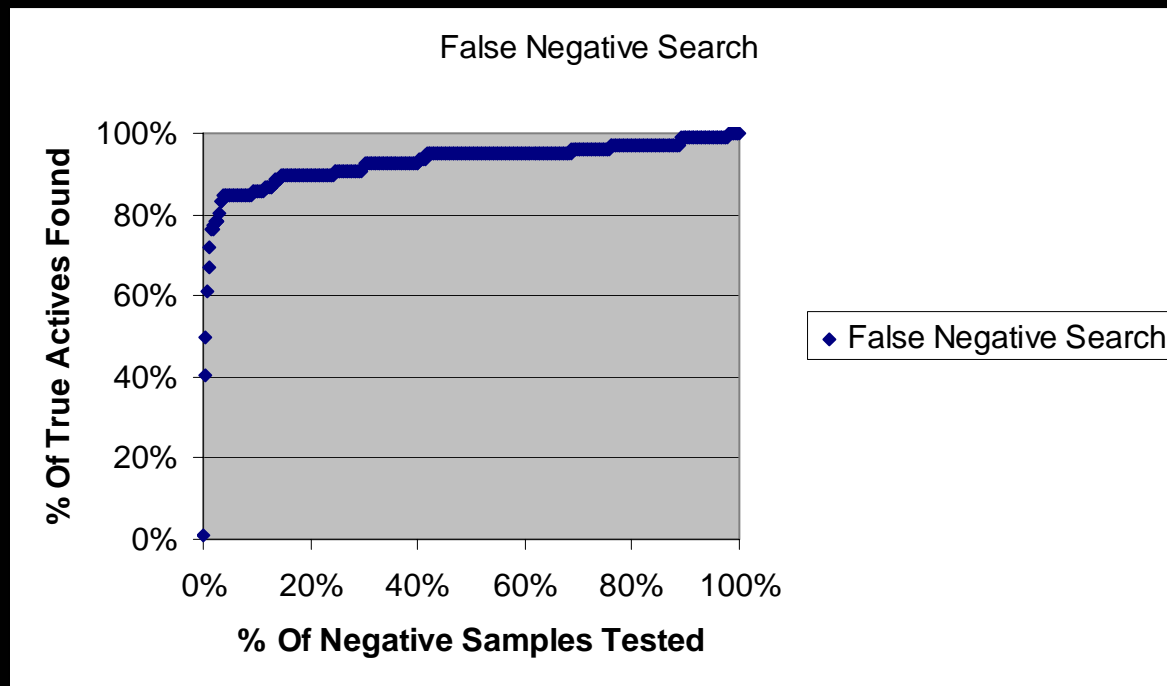
Search for false negatives

- 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

- Introduction to data pipelining
- 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 (each screening point costs US\$0.25 upwards)

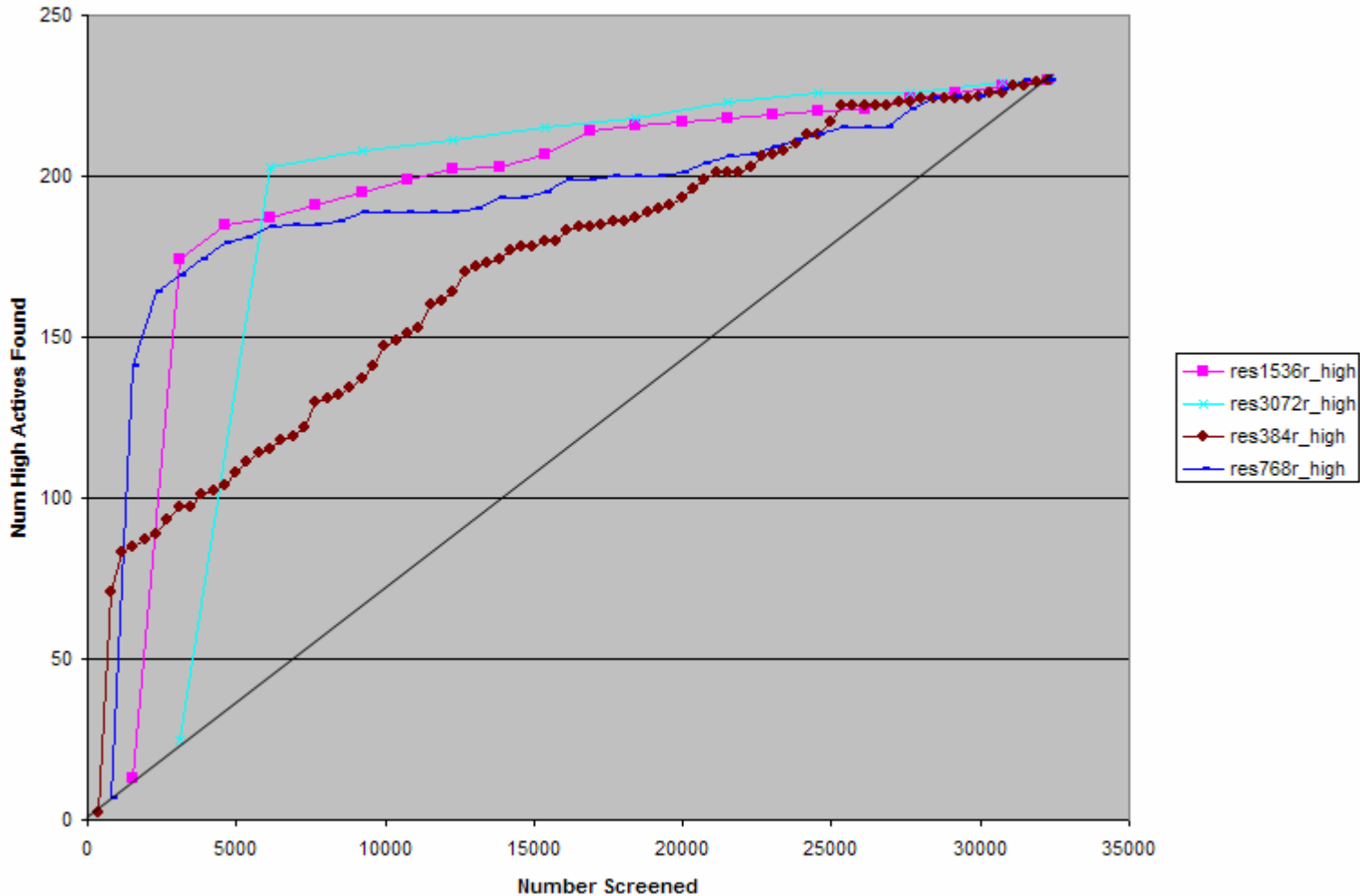
- Iterative screening
 - Screen a subset
 - Random / Ordered
 - Build a model of the screening results
 - Prioritize the remaining compounds and select the next subset to screen
 - Update the model and select the next subset
 - Repeat until
 - No more compounds
 - Hit rate falls below a set level

Example

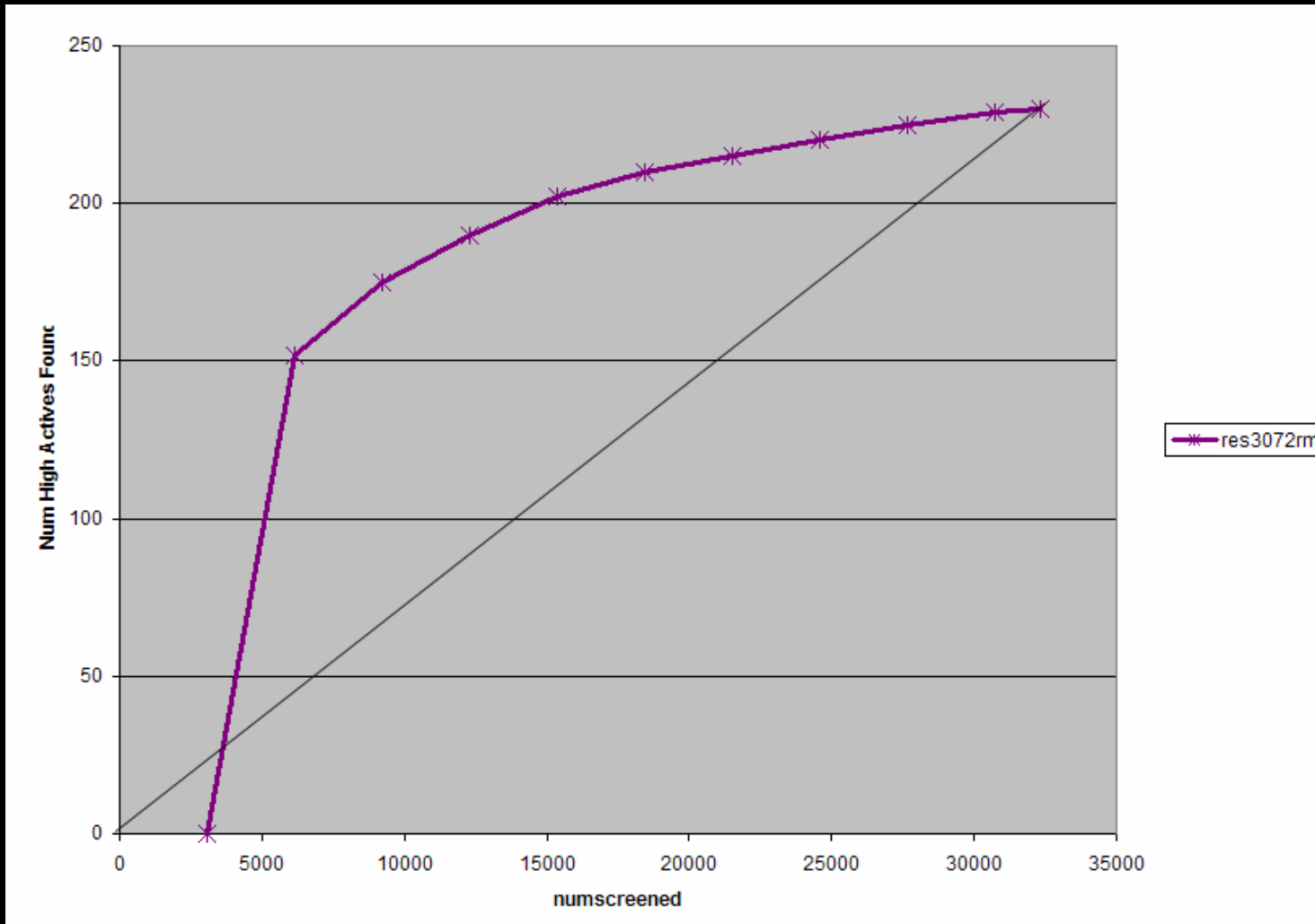
- Using the same NCI AIDS data set
 - Select a subset at random (384, 768, 1536, 3072)
 - “Screen” (i.e look up # actives)
 - Build a Bayesian model
 - Score the remaining compounds
 - Sort by score
 - Select the next subset of the same size and “screen”
 - Repeat until all molecules are “screened”

- Additional experiment
 - Restrict the initial random subset to weakly actives

Initial set contains CA (Actives)

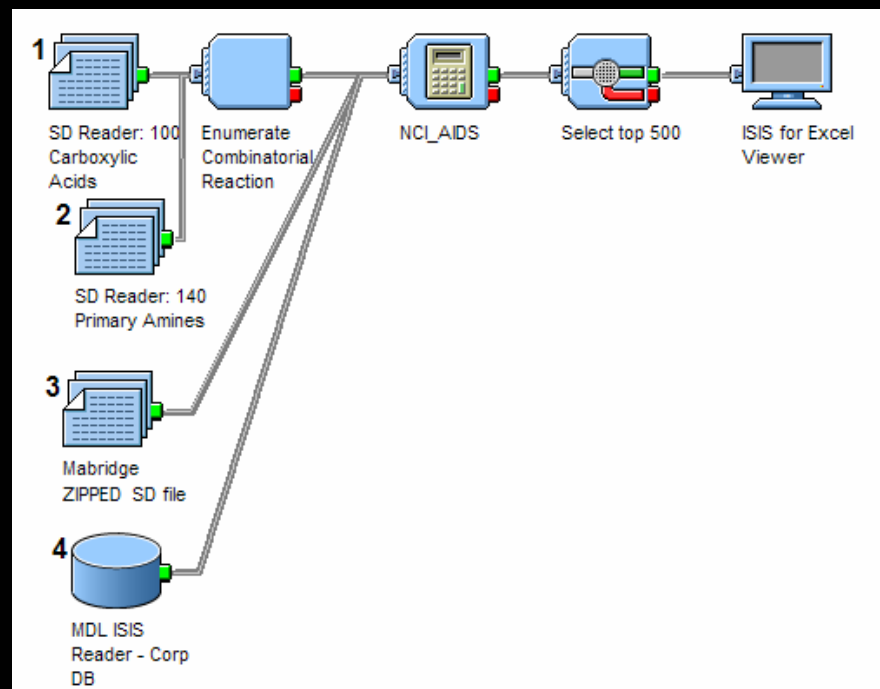


Initial set contains only CM (Weak Actives)



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



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
- Data pipelining provides the infrastructure for successful deployment of virtual screening