

Multicriteria Modeling: The Next Stage in Handling Large Data Sets

David Rogers

SciTegic, Inc.

Anaheim ACS Meeting 2004



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

Goals

- Modeling Large Amounts of Data
 - Original goal of Pipeline Pilot
 - Fast, approximate “non-fitted” models
- Comparing activity classes
 - May be related
 - *agonist vs antagonist*
 - May be quite different
 - often seen as “side effects”
 - May be a mixture of activities and properties
 - solubility, bioavailability, etc.

Sections of Talk

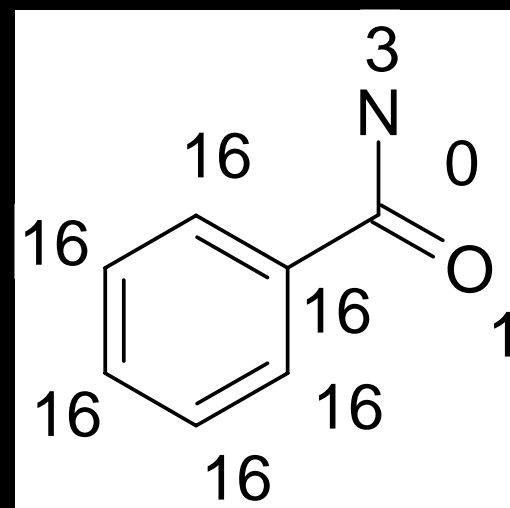
- Extended-connectivity fingerprints
- Modified Bayesian Modeling
- Learning 5HT drugs
- Using multiple models to study compounds
 - 5HT-1A agonists, antagonists, reuptake inhibitors
- Studying features shared by models
 - 5HT-1A agonism vs antagonism features
 - Antagonism vs solubility
- Conclusions

ECFP: Extended Connectivity Fingerprints

- Bits are derived from intermediate results within a *Morgan extended-connectivity* calculation
- Each bit represents a structural (not substructural) feature
- 4 Billion different bits
- Multiple levels of abstraction contained in single FP
- Different starting atom codes give different FPs (ECFP, FCFP, ...)
- Typical molecule generates 100s - 1000s of bits
- Typical library generates 100K - 10M different bits.

Initial Atom Codes

- For ECFPs, we use the Daylight invariant code (atom type, charge, numH, connectivity)
- For FCFPs, the code is created by combining bits representing the functional roles played by the atom:
 - 1: Has lone pairs
 - 2: Is H-bond donor
 - 4: Is negative ionizable
 - 8: Is positive ionizable
 - 16: Is aromatic
 - 32: Is halogen

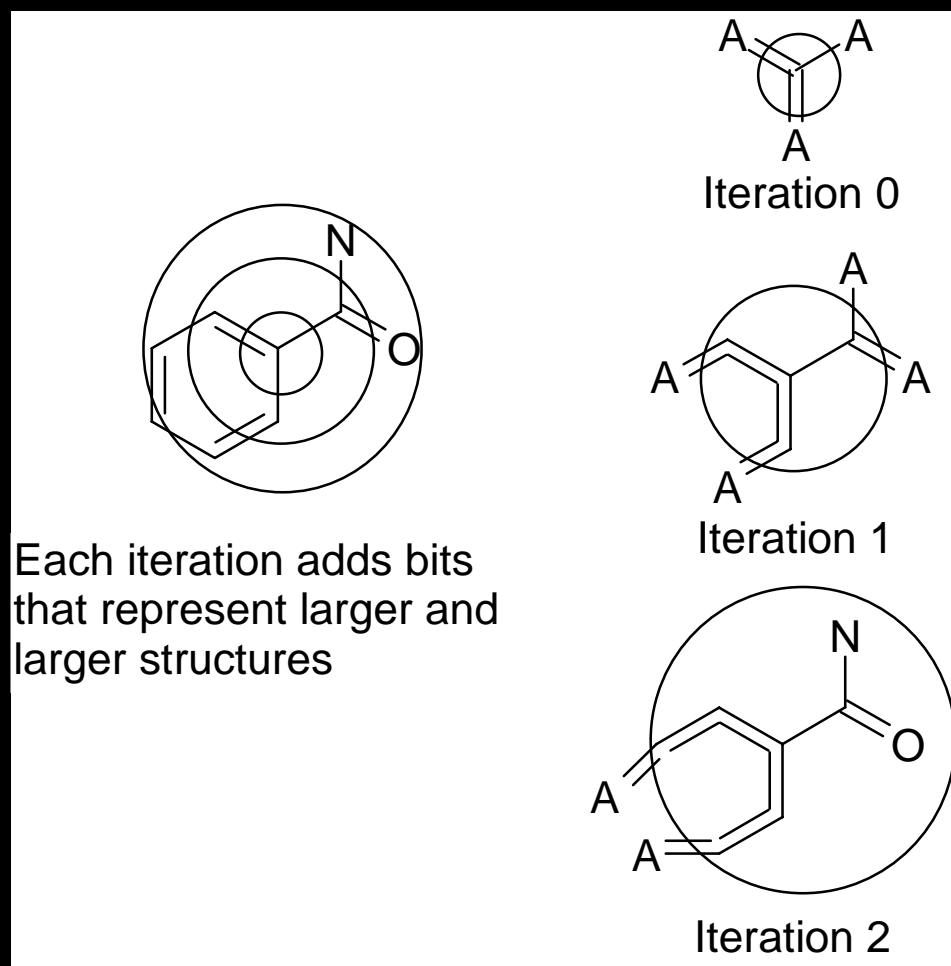


ECFP: Extending the Initial Atom Codes

At each iteration, the code of an atom is “hashed” together with the codes of its neighbors

The code represents a structure captured by a circular “cookie-cutter”

Each atom is used as a center for multiple, different diameter, cuts



ECFP: Generating the Fingerprint

- Iteration is repeated desired number of times
 - Each iteration extends the diameter of the cutter by two bonds
- Codes from all iterations are collected
- Duplicate bits may be removed

```
> <FCFP_0#S>
16
0
1
3
...
```

```
> <FCFP_2#S>
16
0
1
3
1618154665
203677720
-1549103449
1872154524
1070061035
...
```

```
> <FCFP_4#S>
16
0
1
3
1618154665
203677720
-1549103449
1872154524
1070061035
991735244
-453677277
-581879738
-1094243697
690083042
-975279903
...
```

Bayesian Learning

- Build a model to separate “good” from “bad”
- We use Lapacian-modified Naïve Bayesian statistics
 - Efficient:
 - scales linearly with large data sets; single-pass
 - No “fitting” step
 - Robust:
 - works for a *few* as well as *many* ‘good’ examples (“skewed” data)
 - no tuning parameters needed
 - Extrapolation asymptotes to don’t-know
 - Multimodal:
 - can model broad classes of compounds
 - multiple modes of action represented in a single model

Laplacian Weighting of Features

- The *Laplacian probability* $P(\text{act}|\text{feature})$ is used because different features are sampled different numbers of times.
- Given the number of samples containing a particular feature is F_{total} , and the number of those samples that are active is F_{active} .
- If we had a sufficiently-large number of samples of that feature, we might use the estimate $P(\text{act}|\text{feature}) = F_{\text{active}} / F_{\text{total}}$.
- However, if the number of samples is small, the odds are that the estimate is biased by sampling error
- The Laplacian correction can be thought of as adding some number N of “virtual” samples to the counts, at baseline probability. The corrected estimator is $P(\text{act}|\text{feature}) = (F_{\text{active}} + N * P(\text{act})) / (F_{\text{total}} + N)$.

Assumptions

- Features are independent
 - Decorrelation preprocessing not needed
 - Molecule is scored by adding scores of individual features
 - Keeps results “interpretable”
- Collision rate in hash space is low
 - Rate is about a hundred for 100,000’s of features
- Negative information not used
 - Molecules “don’t contain” lots of features
 - Non-Bayesian assumption
 - Negative information from features a sample *does* contain is used
- Result is a unitless number
 - Most sample sets are too biased for “absolute” prediction

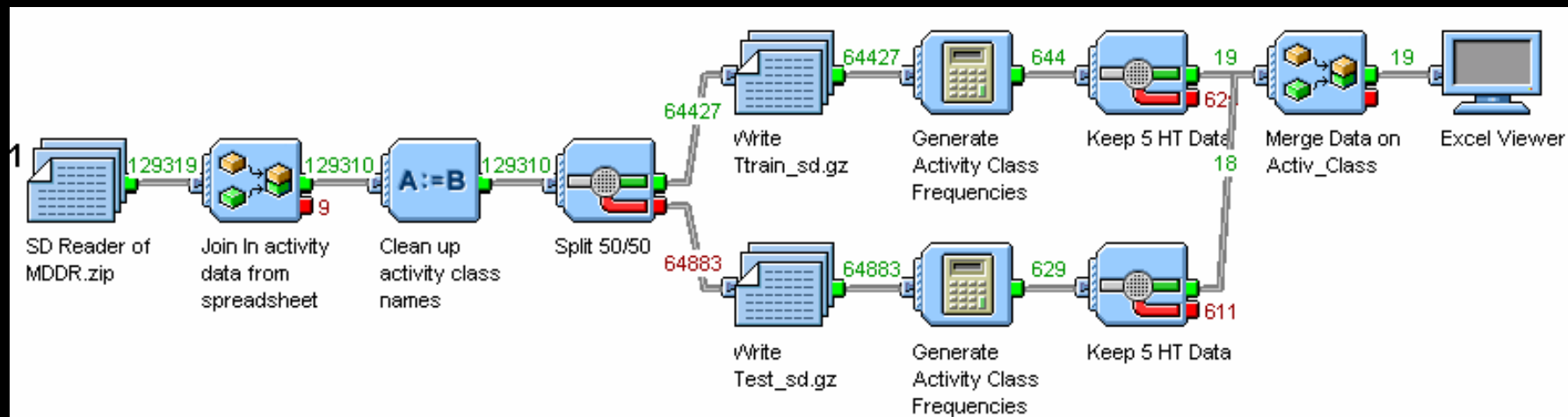
Bayesian References

- Labute, P., *Binary QSAR: A New Method for Quantitative Structure Activity Relationships*; Proceedings of the 1999 Pacific Symposium; World Scientific Publishing, Singapore, 1999.
- Rogers, D., "A Laplacian best-feature model for thrombin inhibitors", in *Designing Drug and Crop Protectants: processes, problems, and solutions*, 2003, Blackwell Publishing, Malden, Massachusetts, eds. Ford, M., Livingston, D., Dearden, J., Van de Waterbeemd, H.
- Bender, A., Mussa, H., Glen, R., Reiling, S., *Molecular Similarity Searching Using Atom Environments, Information-Based Feature Selection, and a Naive Bayesian Classifier*, J. Chem. Inf. Comp. Sci., 44, 2004, pp. 170-178

5-Hydroxytryptamine (5 HT) Receptors

- Receptors for serotonin
 - Mostly target G-protein coupled receptors
 - (5 HT₃ regulates a ligand-operated ion channel)
- Widespread throughout the body
- Large number of receptor subtypes
 - HT₁₋₇, with further subsubtypes (e.g., 5 HT_{1A})
- Important therapeutic effects
 - Antidepressants (5 HT Reuptake Inhibitors)
 - Antinausea (Antagonists of 5 HT₃)
 - Antipsychotics (Antagonists of 5 HT₂)
 - Migraine (Antagonists of 5 HT_{1D}, 5 HT_{2A}, 5 HT_{2C})

5 HT Activity Classes

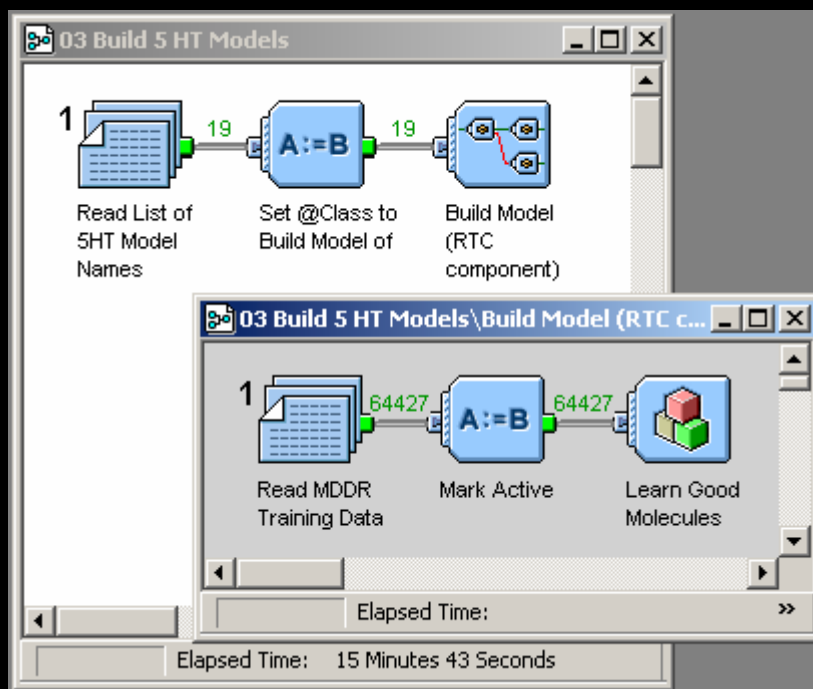


- 19 different 5 HT classes in MDDR (some molecules in more than one class)
- Randomly split into training and test data sets

ACTIVITY CLASS	#TRAINING	#TEST
5 HT Antagonist	29	31
5 HT Reuptake Inhibitor	298	281
5 HT1 Agonist	33	40
5 HT1A Agonist	435	479
5 HT1A Antagonist	221	238
5 HT1B Agonist	28	22
5 HT1C Agonist	47	55
5 HT1C Antagonist	4	0
5 HT1D Agonist	309	305
5 HT1D Antagonist	153	151
5 HT1F Agonist	47	42
5 HT2 Antagonist	84	72
5 HT2A Antagonist	245	263
5 HT2B Antagonist	48	45
5 HT2C Antagonist	128	100
5 HT3 Agonist	28	25
5 HT3 Antagonist	372	463
5 HT4 Agonist	113	93
5 HT4 Antagonist	91	126

Automatically Building 19 5 HT Models

- “Run To Completion” subprotocol used to build a model for each 5 HT activity class



Name	Frequency
5 HT1C Agonist_Model	4236
5 HT1D Antagonist_Model	15012
5 HT1A Agonist_Model	33996
5 HT2C Antagonist_Model	11871
5 HT4 Antagonist_Model	9313
5 HT4 Agonist_Model	11089
5 HT1C Antagonist_Model	496
5 HT1F Agonist_Model	5170
5 HT1B Agonist_Model	3681
5 HT1D Agonist_Model	25540
5 HT2B Antagonist_Model	4983
5 HT3 Antagonist_Model	28383
5 HT2A Antagonist_Model	20847
5 HT3 Agonist_Model	2557
5 HT Reuptake Inhibitor_Model	23858
5 HT Antagonist_Model	3420
5 HT1A Antagonist_Model	19584
5 HT2 Antagonist_Model	9004
5 HT1 Agonist_Model	3570
Total Features	239614
Total Unique Features	33996

Applying the models to the test data

- Models applied against the test data
- ROC scores used to judge model quality

Activ_Class	#Train	#Test	ROCscore
5 HT1C Agonist	50	52	0.999
5 HT4 Antagonist	109	108	0.999
5 HT1F Agonist	40	49	0.999
5 HT Antagonist	38	22	0.998
5 HT4 Agonist	92	114	0.994
5 HT1D Antagonist	161	142	0.996
5 HT3 Agonist	26	27	0.993
5 HT1D Agonist	307	305	0.998
5 HT3 Antagonist	433	401	0.996
5 HT1 Agonist	36	37	0.982
5 HT2B Antagonist	44	49	0.99
5 HT1B Agonist	24	26	0.995
5 HT2C Antagonist	114	114	0.968
5 HT2 Antagonist	69	86	0.988
5 HT1A Agonist	445	469	0.99
5 HT1A Antagonist	219	240	0.989
5 HT2A Antagonist	245	263	0.977
5 HT Reuptake Inhibitor	285	294	0.986

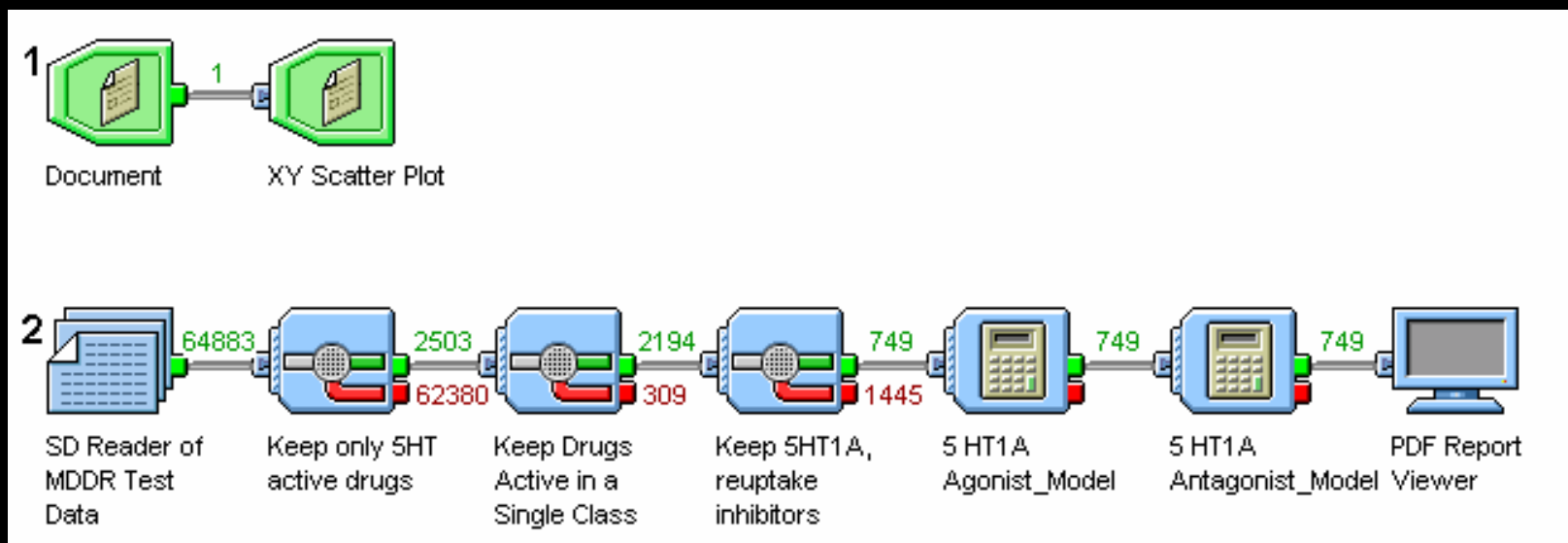
Models build using only 5 HT

- Concern that all models are the same:
 - “They all distinguish 5 HT drugs from all other drugs”
 - “Might not be able to detect subtle differences among 5 HT drugs themselves”
- Repeat model building and testing using only 5HT-active compounds

Activ_Class	#Train	#Test	ROCscore	ROCscore_5HT
5 HT1C Agonist	50	52	0.999	0.997
5 HT4 Antagonist	109	108	0.999	0.992
5 HT1F Agonist	40	49	0.999	0.984
5 HT Antagonist	38	22	0.998	0.983
5 HT4 Agonist	92	114	0.994	0.983
5 HT1D Antagonist	161	142	0.996	0.979
5 HT3 Agonist	26	27	0.993	0.977
5 HT1D Agonist	307	305	0.998	0.976
5 HT3 Antagonist	433	401	0.996	0.975
5 HT1 Agonist	36	37	0.982	0.969
5 HT2B Antagonist	44	49	0.99	0.968
5 HT1B Agonist	24	26	0.995	0.957
5 HT2C Antagonist	114	114	0.968	0.942
5 HT2 Antagonist	69	86	0.988	0.937
5 HT1A Agonist	445	469	0.99	0.922
5 HT1A Antagonist	219	240	0.989	0.922
5 HT2A Antagonist	245	263	0.977	0.899
5 HT Reuptake Inhibitor	285	294	0.986	0.891

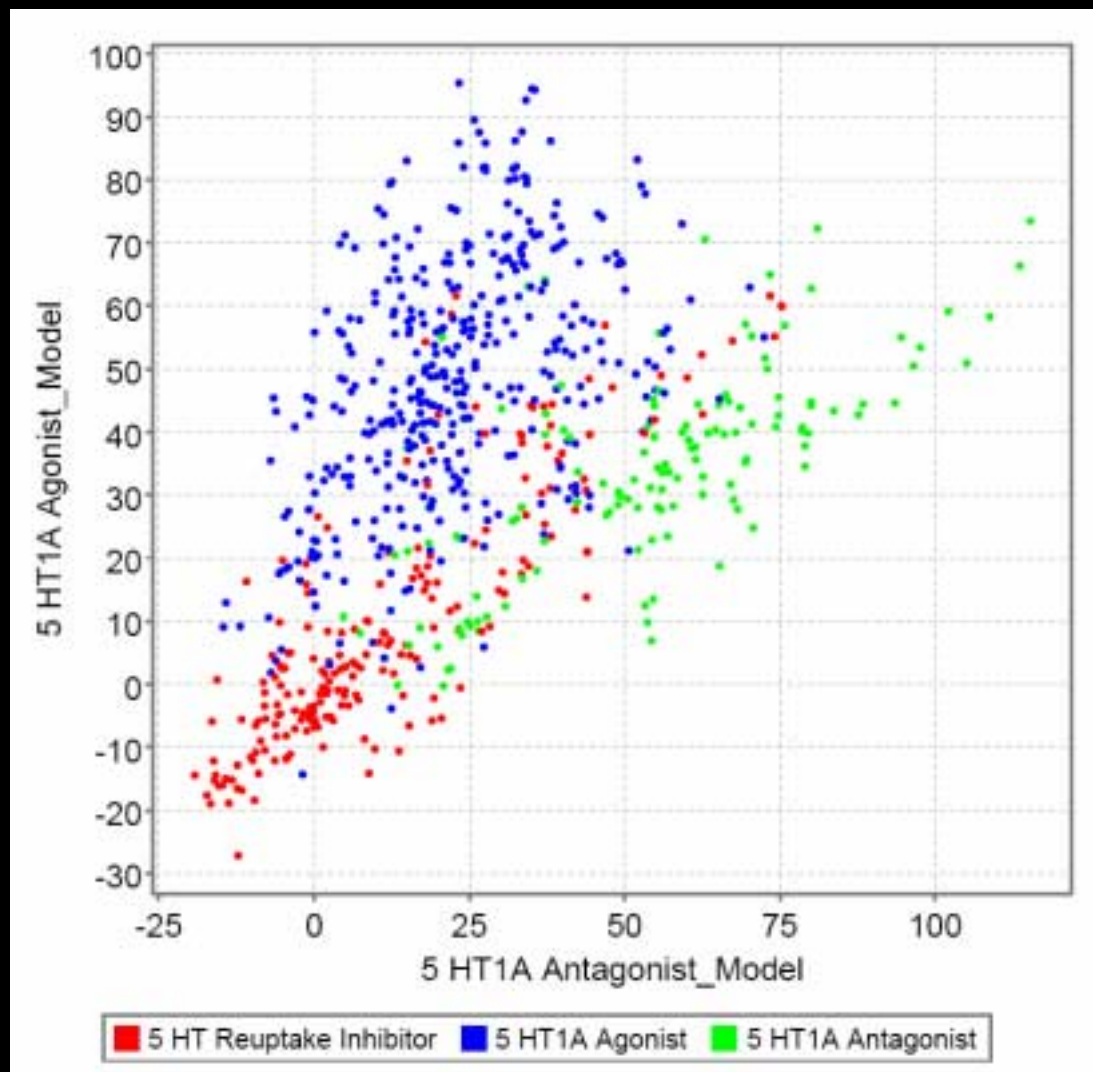
Using Multiple Models

- Multiple models can be used to cluster, visualize, or suggest potential side effects
- In this example, we view an XY scatter plot of:
 - 5 HT_{1A} agonists, 5 HT_{1A} antagonists, and 5 HT reuptake inhibitors
 - Using the 5 HT_{1A} Agonist Model vs. the 5 HT_{1A} Antagonist Model



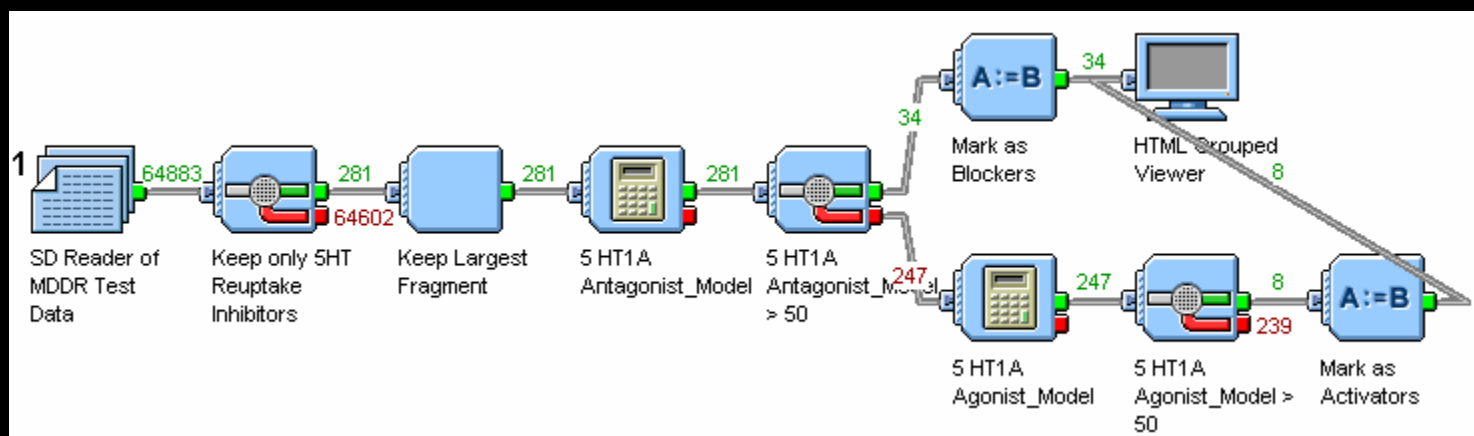
Results

- (All data from the test set)
- 5 HT_{1A} agonists and antagonists mostly separated
- Note some 5HT reuptake inhibitors “look like” 5HT_{1A} agonists



Viewing blockers and activators

- The two groups of 5HT_{1A}-active reuptake inhibitors identified in the scatterplot can be viewed
 - Blockers*: Molecules that score high (>50) using the antagonist model
 - Activators*: Molecules that score high (>50) using the agonist model, but lower (<50) using the antagonist model



Reuptake inhibitors interacting with 5 HT_{1A}

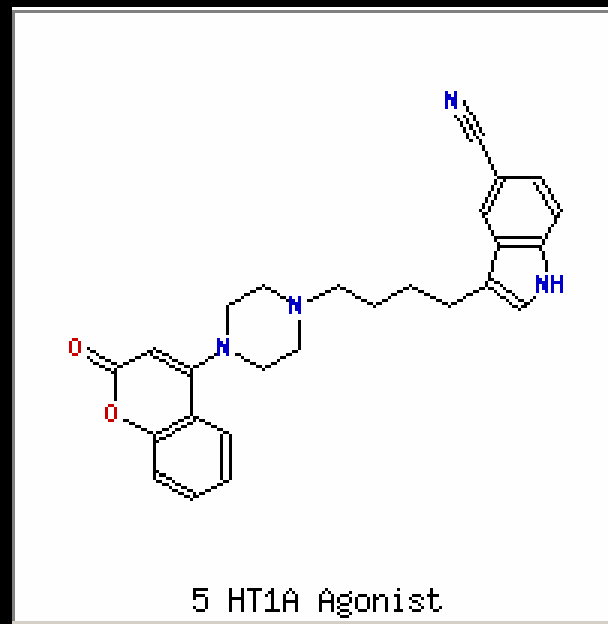
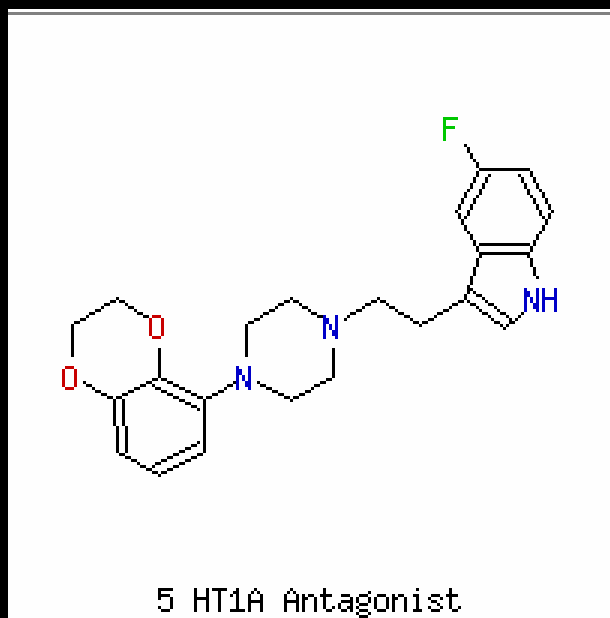
- Caption shows if a reuptake inhibitor was marked as also having 5HT_{1A} agonist or antagonist activity

Role: Blocker		
<p>Chiral</p> <p>5 HT1A Antagonist</p>	<p>5 HT1A Antagonist</p>	<p>5 HT1A Antagonist</p>
<p>5 HT1A Antagonist</p>	<p>5 HT1A Antagonist</p>	<p>5 HT1A Antagonist</p>
<p>5 HT1A Antagonist</p>	<p>5 HT1A Antagonist</p>	<p>5 HT1A Antagonist</p>

Role: Activator		
<p>5 HT1A Agonist</p>	<p>5 HT1A Agonist</p>	<p>5 HT1A Agonist</p>
<p>5 HT1A Agonist</p>	<p>5 HT1A Agonist</p>	<p>5 HT1A Agonist</p>
<p>5 HT1A Agonist</p>	<p>5 HT1A Agonist</p>	<p>5 HT1A Agonist</p>

Closeup of one blocker and one activator

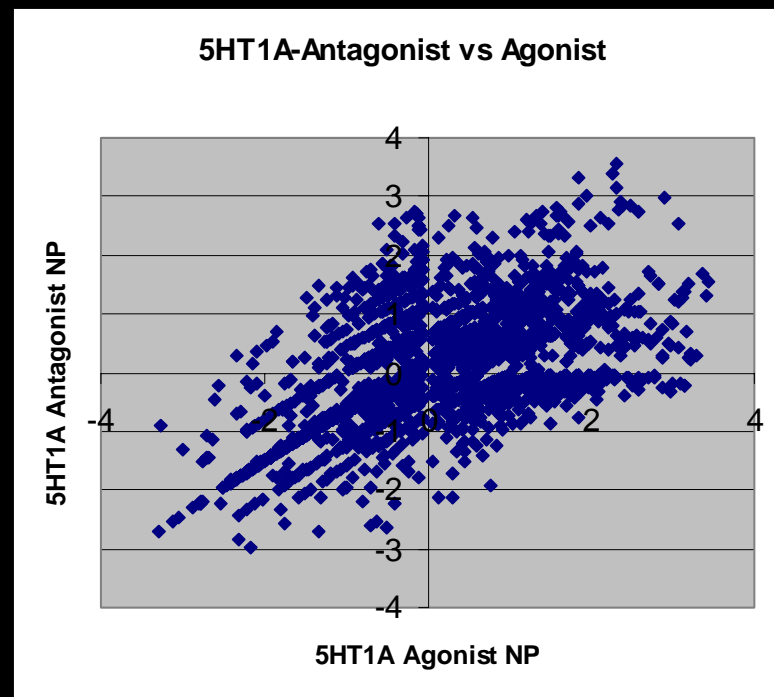
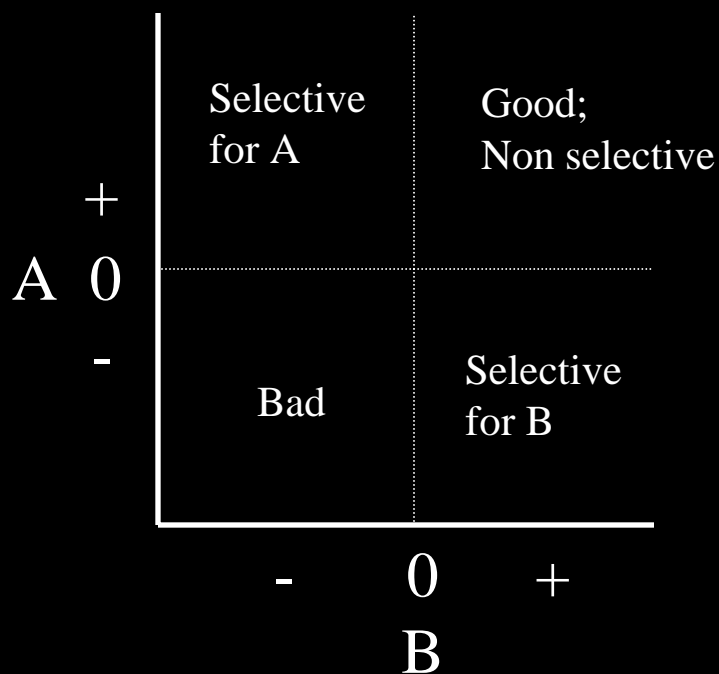
- Zooming in on one of the blockers and one of the activators:



- Despite their similarity, both molecules were correctly assigned to the correct class, as reflected in alternate activity class info.

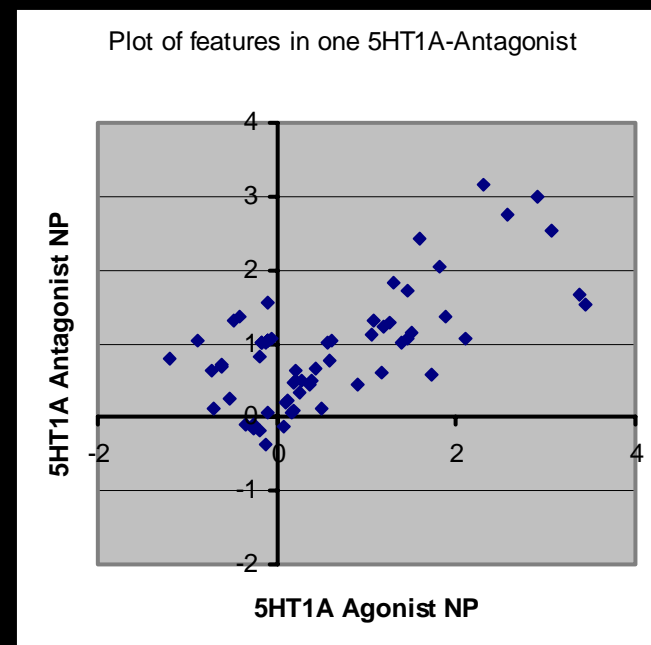
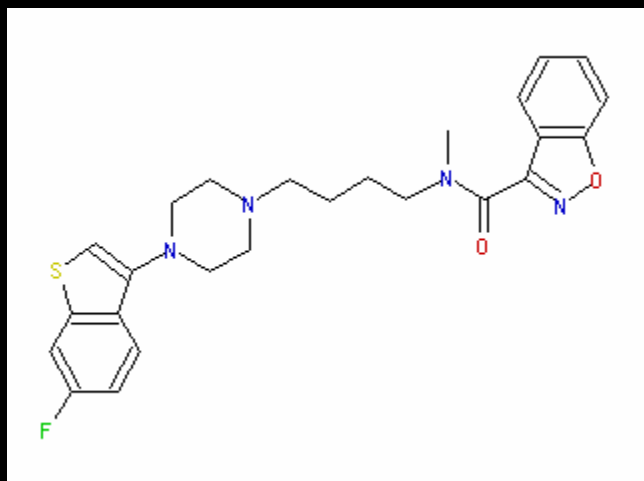
Design for specificity

- 5HT_{1A} agonists and antagonists



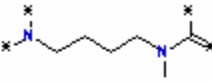
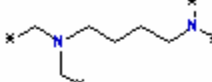
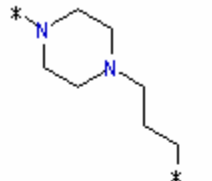
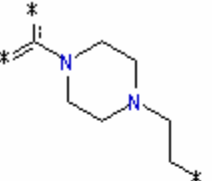
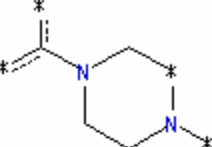
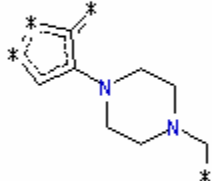
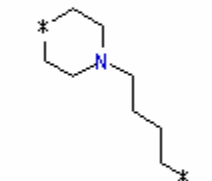
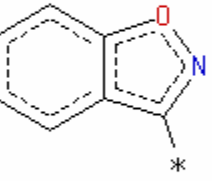
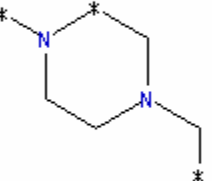
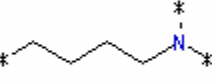
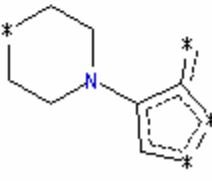
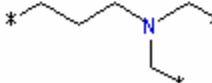
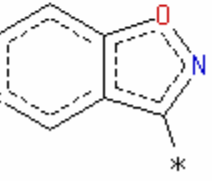
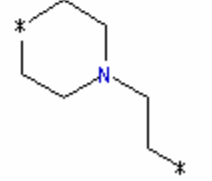
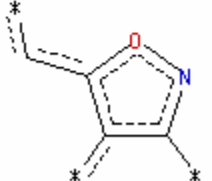
Viewing a single molecule

- Similar scatterplots are available for single compounds
 - E.g., a 5 HT_{1A} antagonist



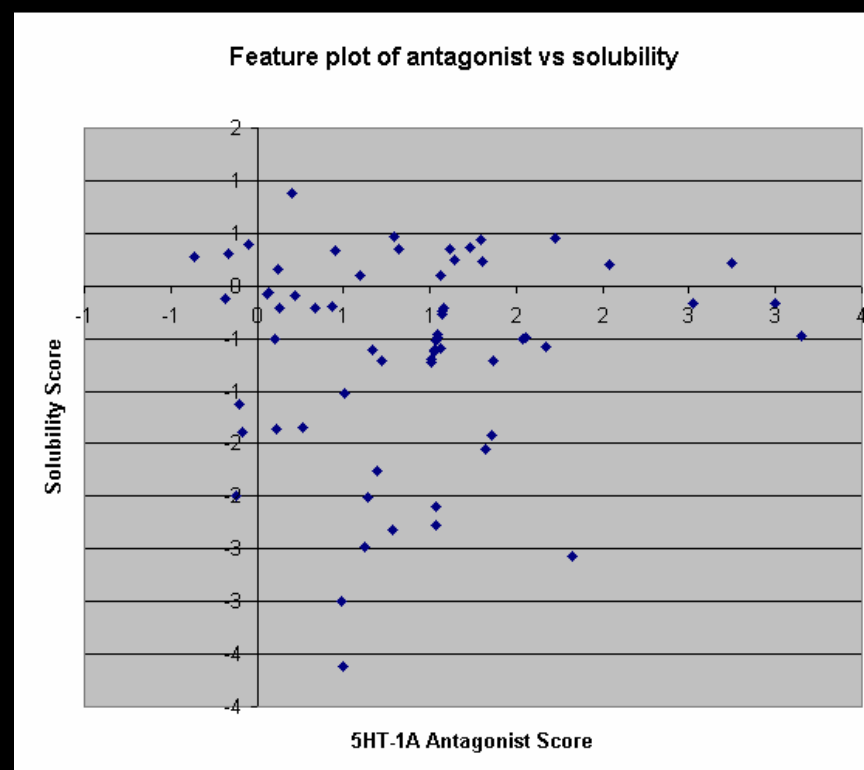
Most agonistic features

- Sort the features by the difference between the agonist and antagonist scores
- The top 24 agonists features:

Group: molecule features				
 Ag 3.436 Antag 1.536	 Ag 3.368 Antag 1.671	 Ag 3.061 Antag 2.521	 Ag 2.911 Antag 2.995	 Ag 2.575 Antag 2.747
 Ag 2.307 Antag 3.149	 Ag 2.097 Antag 1.060	 Ag 1.892 Antag 1.365	 Ag 1.822 Antag 2.037	 Ag 1.735 Antag 0.591
 Ag 1.593 Antag 2.438	 Ag 1.497 Antag 1.145	 Ag 1.458 Antag 1.065	 Ag 1.456 Antag 1.729	 Ag 1.384 Antag 1.011

Antagonism vs Solubility

- By building a model of solubility, the effect of features on solubility vs. antagonism can be compared
- Results can suggest alterations to effect one criteria with minimal effect on another



Conclusions

- Handling large amounts of data “not about a single target anymore”
- Possible next steps:
 - A single model with 100’s or 1000’s of categories
 - Renormalization of scores for “side-effect” predictor
 - “Continuous categories”
 - Better visualizations of results
- Thanks to:
 - Elie Giraud, Aventis, for the inspiration
 - Phil Cochrane (Java Scatterplot Viewer)