

# **Chemical Informatics using JMP™ PowerMV and ChemModLab**

**S. Stanley Young**, National Institute of Statistical Sciences

**Thomas H. Burger**, Research Consultant

Michael S. Lajiness, Eli Lilly and Company

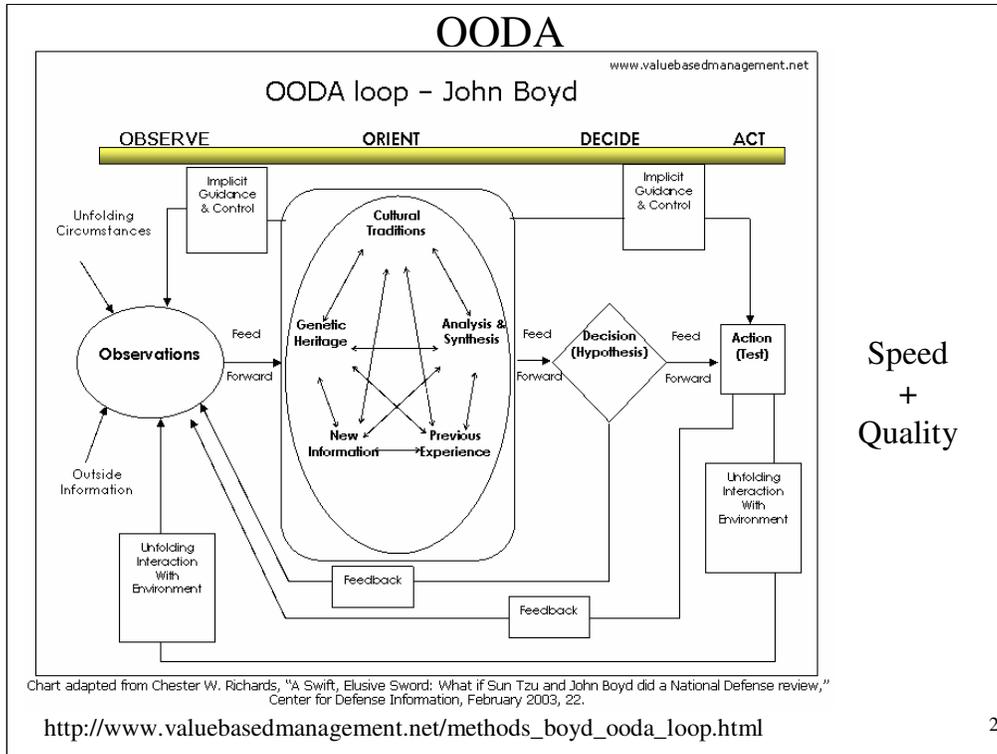
Wolf-D. Ihlenfeldt, Xemistry

JMP Users' Conference

Oct 12-14, 2007

1

Drug discovery strategies use various software tools. There is a lot of specialized commercial software, but it is typically quite expensive. In this lecture, the use of several software tools are presented that can be used in combination with SAS JMP.



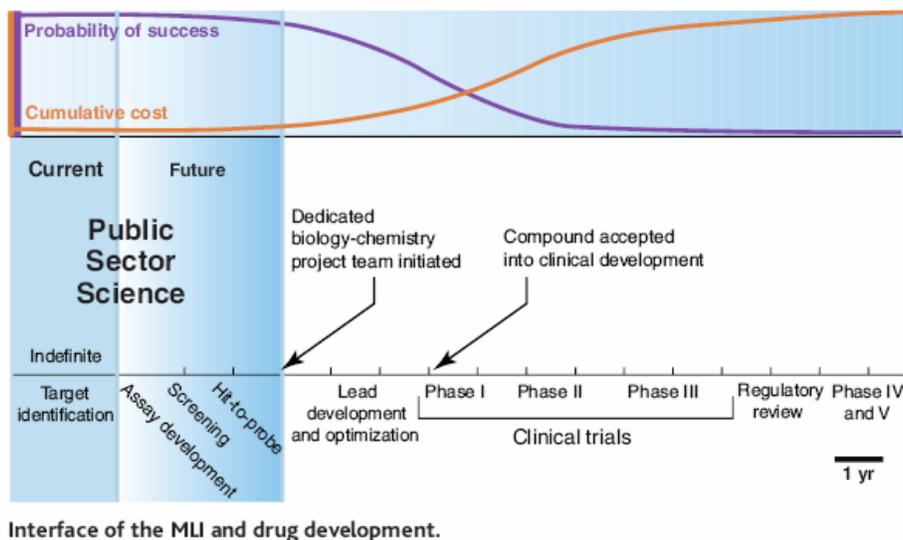
John Boyd spoke of the cycle time in combat. Cycle time is critical in a multiple cycle process. The basic idea is to understand the situation quickly and respond faster than your opponent.

There are some real problems in that the FDA controls a critical decision point and they are not very fast. Drug development could be a lot faster in Japan or China, for example, if they put compounds into humans faster. Arguably the world would be better off if drugs came onto the market faster and if problems are identified, come off the market faster as well.

We present several tools to help understand complex data and make decisions quicker.

You get the philosophy for free.

# Drug Discovery / Development Pipe Line



3

There are huge variation in cycle time. Also problems in transition from one step in the process to the next. The point is to assess the current situation quickly and respond to move the discovery process along.

# Motivation

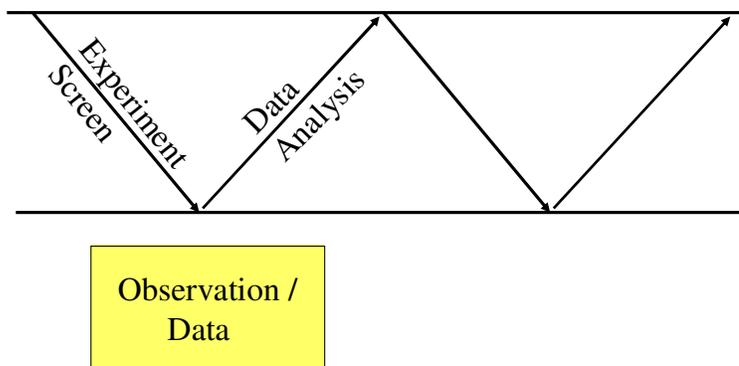
1. Drug discovery is a search through a HighD world.
2. Drug discovery follows the George Box theory/data alternation.
3. Success is based on Data, Analytics, Visualization.
4. Our story is Experimental Data, JMP, PowerMV and ChemModLab.

4

The point is to empower the scientist to understand data more quickly so that good decisions can be made faster than are being made by others.

# George Box, Iterative Discovery

Hypothesis / Theory

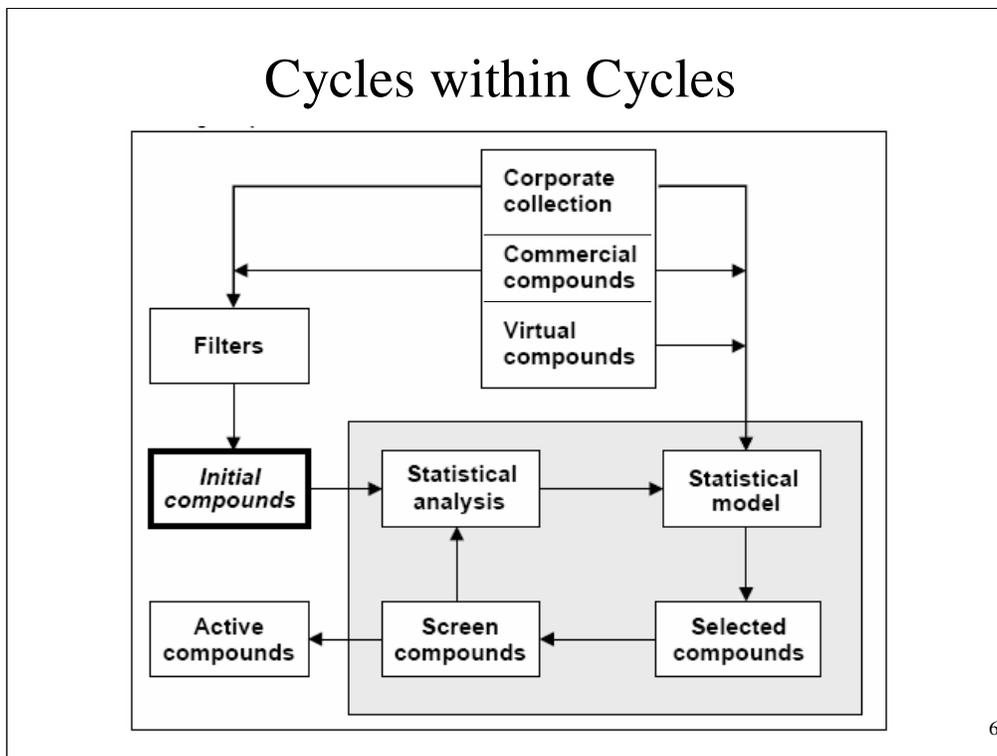


5

There is the needle in a haystack search problem. How do you find a good compound among millions of compounds?

We have access to very large sets of compounds. Some of these compounds have been tested. We need to understand the tested compounds so that we can select additional compounds for testing. We develop some theory about the important features of active compounds. We select compounds for screening and use the new data to build a better model of what makes compounds active.

## Cycles within Cycles



All drug discovery/development is cycles and cycles within cycles. How fast can we complete the cycles?

Deep in research the value of one day saved is \$80,000. Time is money.

Sequential screening proceeds as follows:

1. A training set is constructed from the corporate collection using various filters.
2. The compounds are biologically tested and the results subjected to statistical analysis.
3. The resulting model of important features is used to select additional compounds for screening.
4. Steps 2 and 3 are repeated, giving rise to active compounds.

## Structure Activity Relationship

"Structure-Activity Relationship" (SAR) –  
in practice is finding the experimental  
relationship between the structures and  
biological activity for a series of compounds.

7

In short, SAR is to find the relationship between structures and their activities based on statistical analysis.

# Components of SAR Analysis

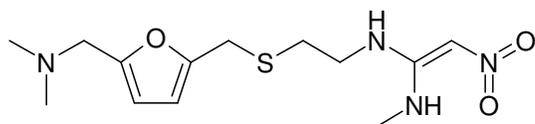
1. Target properties (discrete or continuous)  
(the “y” of regression)  
Activity data from various bioassays (EC50, IC50, 0/1)
2. Descriptors (the “X” of regression)
  - a. Based on chemical formula (or 1D)
  - b. Based on chemical graph (topological, or 2D)
  - c. Based on molecular conformations (3D)
3. Statistical Analysis
  - a. Recursive Partitioning
  - b. Cluster Analysis
  - c. Classification/Regression Analysis (Support Vector Machine, Random Forest, etc. )

8

Statistical analysis can only accept numbers; it can not understand compound structures, therefore, special algorithms are used to compute numbers to represent structures.

Once the structures are encoded into numbers, statistical analysis can be engaged to reveal the relationship between the activity and structure.

## Descriptor Generation



Graph

### Numbers

Continuous Descriptors: 0.73, 0.30, 0.387, . . .  
Discrete Descriptors: 3, 2, 7, . . .  
Bit String: 01011011 ...

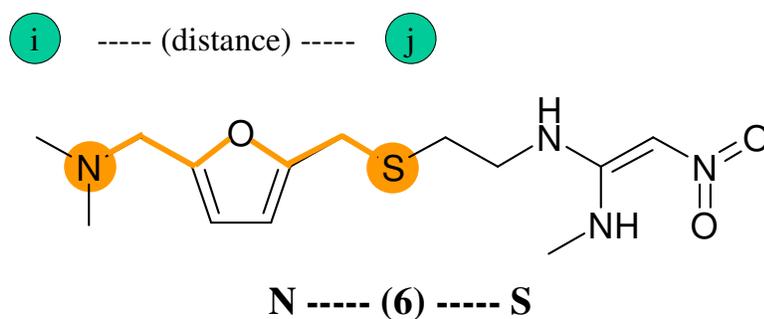
9

Given a chemical structure, three types of descriptors can be generated. Continuous descriptors like BCUTs, discrete descriptors like atom pair count, and bit string descriptors like molecular fingerprints. These descriptors can be used in different aspects of QSAR modeling.

Although there is no agreement on the best descriptors, it is observed that many types of descriptors capture relevant information for statistical modeling.

# Atom Pair Descriptors

Topological Distance: shortest chemical graph distance (number of atoms) between atom i and atom j (2D AP) or physical distance (3D AP)



Carhart, R. E.; Smith, D. H.; *J. Chem. Inf. Comput. Sci.* **1985**, 25, 64-73

10

Here, I will give a simple example on atom pair descriptors. Atom pairs have been successfully used in many drug discovery projects.

## Chemical Feature Definition

1. Based on pharmacophore groups
  - a. Hydrogen bond donor, acceptor
  - b. Aromatic ring center, hydrophobic center
  - c. Positive charge center, negative charge center
  
2. Based on individual atom types  
C(sp, sp<sup>2</sup>, sp<sup>3</sup>), O(sp<sup>2</sup>, sp<sup>3</sup>), N(sp, sp<sup>2</sup>, sp<sup>3</sup>)
  
3. Based on any fragments
  - a. -O-, -S-
  - b. -CN, -CF<sub>3</sub>
  - c. ....

11

Besides the Carhart definitions of atom features, we can generalize the definitions to generate atom-group pairs.

Many of the most popular molecular descriptors can be computed using free software, PowerMV.

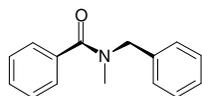
## 2D Bit String Generation

1	0	1	0	1	1	0	1	1	1	0	1	0	0	0	0	0	0	0	0	0	1
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---



ARC\_04\_ARC

Two aromatic ring centers that are 4 bonds away.



12

Here we note the presence of aromatic rings and compute the through-bond distance between them. As chemical bonds are of a rather fixed length, they serve as a good proxy for through-space distance.

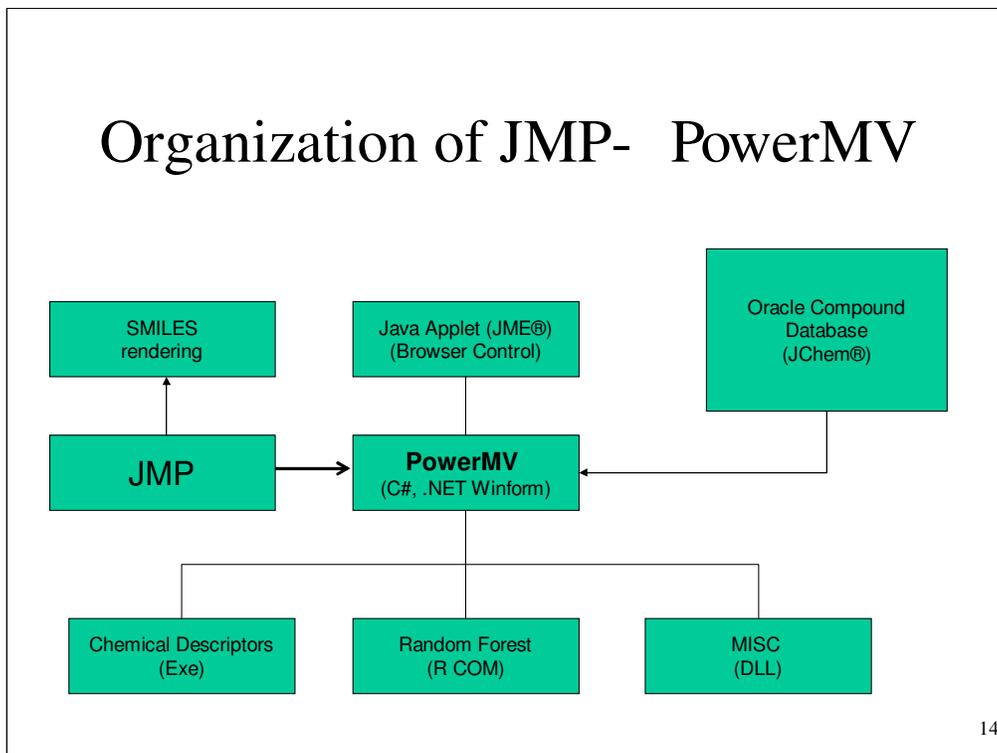
## PowerMV: Major Functions

1. Molecular viewing
2. Report generation
3. Similarity Search & Structure Comparison
4. Chemical descriptor generation
5. Statistical analysis
6. Oracle Database Client

13

PowerMV is a free molecular viewing program written by two post docs, Jack Liu and Jun Feng, when they worked for the National Institute of Statistical Sciences. A public version can be downloaded from [www.niss.org/PowerMV](http://www.niss.org/PowerMV).

## Organization of JMP- PowerMV



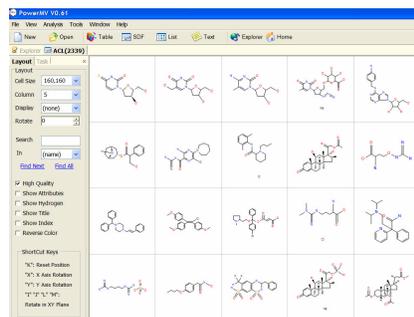
14

PowerMV is a rather complete chemical viewer. You can look at structures. You can compute molecular descriptors. JMP is a very powerful statistical analysis system, but it lacks the ability to view molecules.

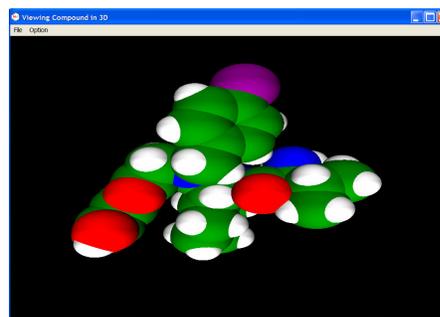
The idea is to compute molecular descriptors in PowerMV and send the descriptors to JMP. Now living in JMP, it is useful to be able to view 2D drawings of molecules. There is a very popular linear character string representation of molecules, SMILES. SMILES strings can be included in the JMP data set.

Our JMP add-on allows a person to convert SMILES strings to 2D drawings. The basic engine for this rendering of a string to a nice drawing was written by Wolf-D. Ihlenfeldt, Xemistry.

# Molecular Viewing



2D

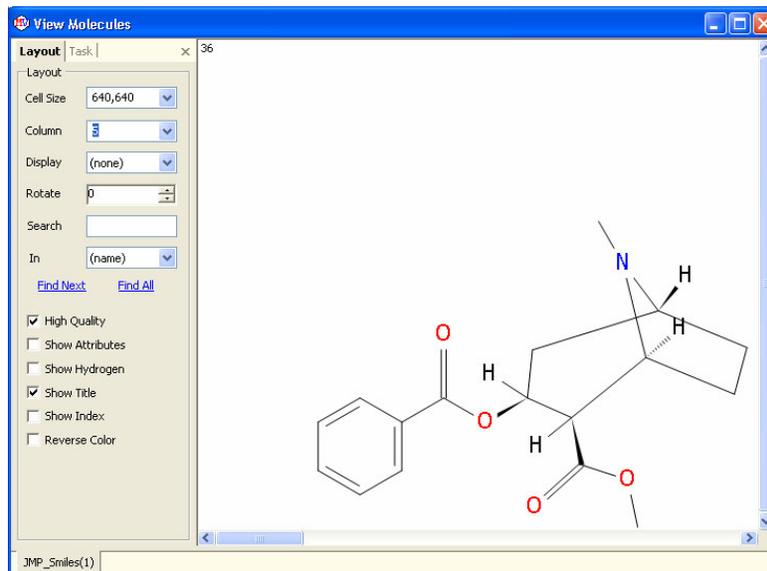


3D

15

Most chemist work from 2D drawings of molecules to do their science. PowerMV has very clean representation of molecules. It allows similarity searching against curated data sets so that the chemist can do mathematical association of the new compound against know compounds.

## View of Compound 36



COC(=O)[C@@H]1[C@@H]2CC[C@H](C[C@@H]1OC(=O)c3ccccc3)N2C

16

Here we have a SMILES string and the 2D drawing of the compound.

The 2D drawing of the molecule is quite useful to the scientist; The smiles string offers many computer science advantages, but easy human comprehension is not one of them.

# Typical Report Generation

The screenshot shows a Microsoft Excel spreadsheet with four rows of data. Each row contains a chemical structure in column A, a name in column B, and a status in column C. Columns D and E contain numerical values. The status 'Unknown' is present in all rows.

	A	B	C	D	E
4		(+/-)-Methoxyverapami	Unknown	0	0.0
5		(+/-)-Normetanephrine	Unknown	0	0.0
6		(+/-)-SKF 38393, N-al	Unknown	0	0.0
		(+/-)-SKF-38393	Unknown	0	0.0

17

Compounds and their features can be exported to Excel. Note the poor use of visual space.

# More effective use of Space

The screenshot shows the PowerMV V0.70 software interface. The main window displays a grid of six chemical structures, numbered 1 through 6. To the left of the grid is a 'Layout' panel with various settings: Cell Size (160,160), Column (3), Display (none), Rotate (0), Search (empty), In (name), Find Next, Find All, High Quality (checked), Show Attributes (checked), Show Hydrogen (unchecked), Show Title (unchecked), Show Index (checked), and Reverse Color (unchecked). Below the grid is a data table with columns: Action, CATNUM, Class, Description, Enzyme, and Name. The table contains six rows of data, each corresponding to a chemical structure in the grid above.

Action	CATNUM	Class	Description	Enzyme	Name
Inhibitor	120693	Neurotransmission	Inhibitor of catecholamine	Enzyme	DL-alpha-Methyl-p
Blocker	144509	Cl- Channel	Cl- channel blocker		N-Phenylanthranil
Inhibitor	190047	Phosphorylation	Potent alkaline phosphatas	Enzyme	S(-)-p-Bromotetra
Antagonist	194336	GABA	Weak GABA-B receptor ant		S-Aminovaleric aci
Inhibitor	211672	GABA	GABA uptake inhibitor		(-)-Nipecotic acid
Inhibitor	246370	DMA Metabolism	Antihistaminic on malona	Enzyme	Asialic acid

18

Much better is to separate the drawings from the numerical and character variables. The pictures and the data table are linked. Click one and the other moves and highlights the corresponding data.

# Chemical Descriptor Generation

Name	Co...	Type
<input type="checkbox"/> Atom Pair	546	<Descriptor>
<input type="checkbox"/> Atom Pair (Carhart)	4662	<Descriptor>
<input type="checkbox"/> Fragment Pair	735	<Descriptor>
<input type="checkbox"/> Pharmacophore Fingerpr...	147	<Descriptor>
<input type="checkbox"/> Weighted Burden Number	24	<Descriptor>
<input type="checkbox"/> Properties	8	<Descriptor>
<input type="checkbox"/> Atom Pair (Carhart) Count	4662	<Descriptor>
<input type="checkbox"/> Fragment Count	90	<Descriptor>
<input type="checkbox"/> EXTREG	1	String
<input type="checkbox"/> LIC50	1	Numeric

1. Generate data sheet directly from SDF files.
2. Calculate 2D atom pair descriptors and eigenvalue based descriptors.
3. Generated data sheet can be directly applied to statistical analysis.

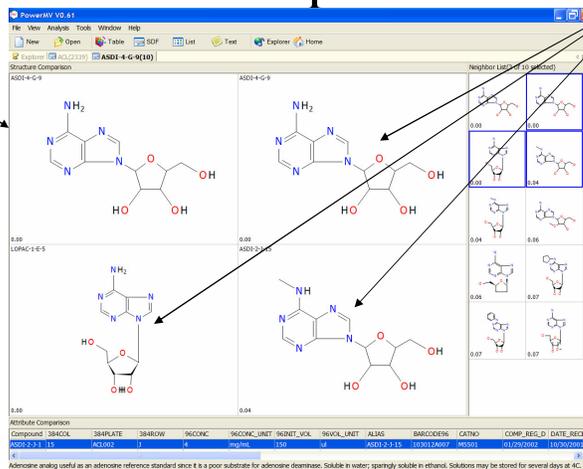
19

Here are the numerical descriptors that can be computed.

# Similarity Search & Structure Comparison

Target

Near Neighbors



20

Similarity searching is an important capability of PowerMV. Near neighbors can be found in a curated data set and the scientist can use the near neighbors to help make judgments about the new compound.

The molecules can be rotated and flipped to align all the compounds in a similar way.

## Measure of Similarity: Tanimoto Distance

Mol\_1 

1	0	1	0	1	1	0	1	1	1	0	1	0	0	0	0	1
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

Mol\_2 

1	1	1	0	1	0	0	0	1	0	0	1	0	0	1	0	0
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

No. of "1" in Mol\_1 AND Mol\_2: 5

No. of "1" in Mol\_1 OR Mol\_2: 11

Tanimoto distance:  $5/11 = 0.455$

21

There has been much research devoted to computing the similarity between molecules.

The first research in similarity seems to have been conducted by a Frenchman, Jaccard. Computational chemist re-invented the method. Basically, only features in common matter. That two compounds have many that are unlike for the two compounds does not make them similar.

# Substructure Search

mol file

```
c2ccc(Nc1ccccc1)cc2
```

JME 2002.05 Thu Apr 19 14:17:55 EDT 2

13 14 0.0000 0.8406 0.0000 V20

Get mol file

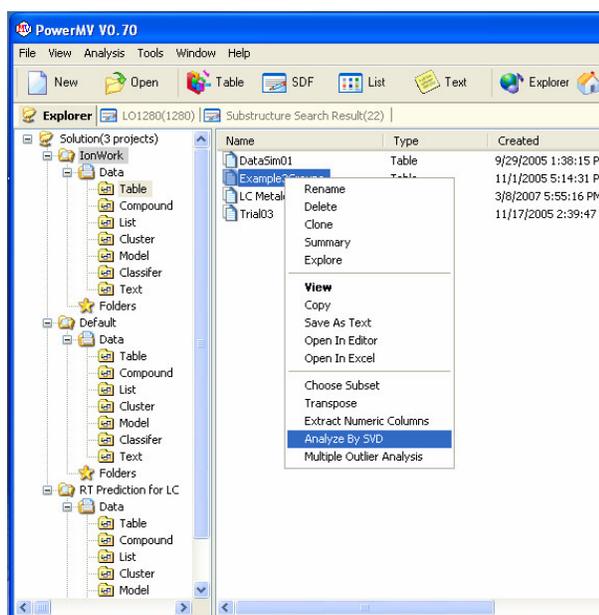
Clear Editor Clear Fields

OK Cancel

22

Often it is useful to do a substructure search. Substructure searching is very computer intensive and difficult. Ad hoc methods are used in combination with graph matching algorithms.

# Unusual Statistical Methods SVD



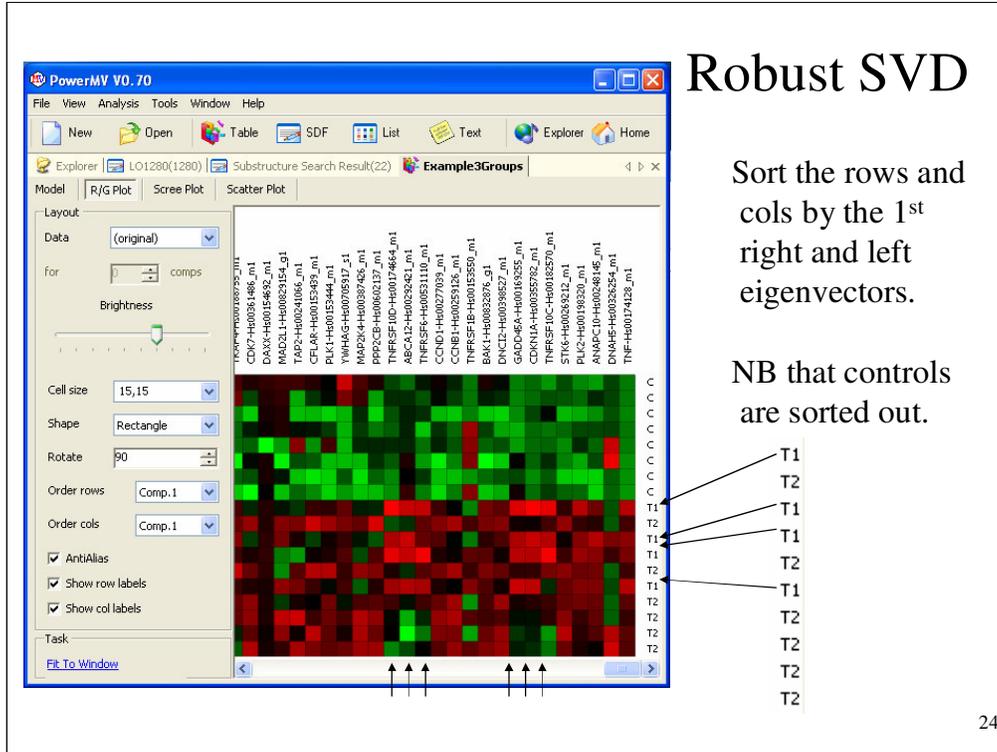
Robust SVD

Outlier detection  
via tetrads

23

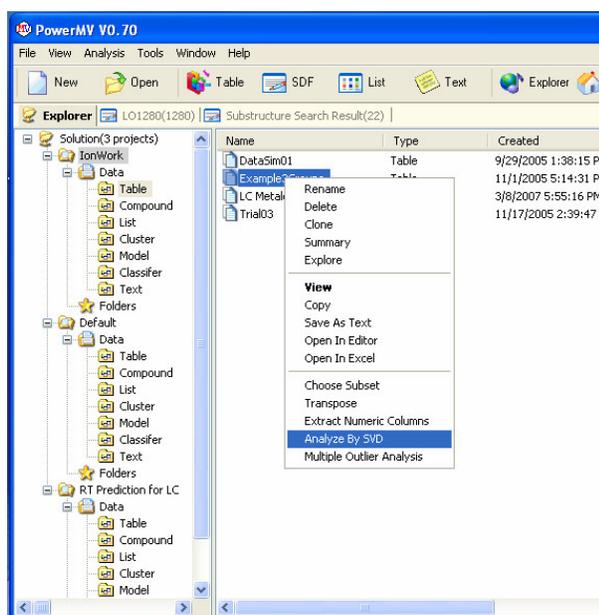
PowerMV has some novel data processing methods. The basis of many statistical methods is singular value decomposition, the factoring of a 2-way table of data. Principle components analysis is the most well-known example of SVD. PowerMV has a version of SVD that is robust to outliers and tolerates a modest amount of missing data. Liu et al. 2003, PNAS.

Outlier detection is quite important in practice. PowerMV has a method by Bradu and Hawkins.



Heat maps of two-way data tables can be sorted by the elements of the right and left eigenvectors of SVD to great advantage. Nice patterns often appear that lead to insight into the nature of the data.

# Unusual Statistical Methods



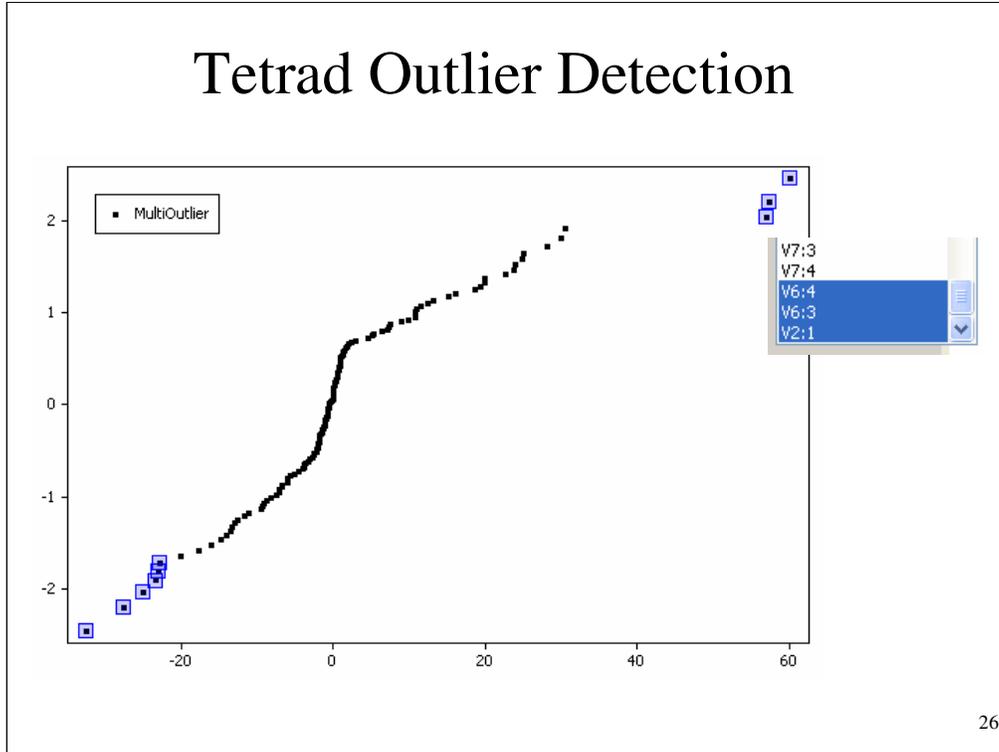
Robust SVD

Outlier detection  
via tetrads

25

Outlier detection is important for two reasons. First, the found outlier might be a bad data point. Many statistical methods are sensitive to bad data. Second, outliers might be correct data points that point to new phenomenon or special interactions.

# Tetrad Outlier Detection



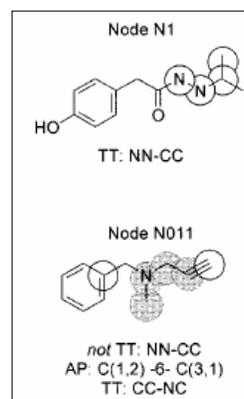
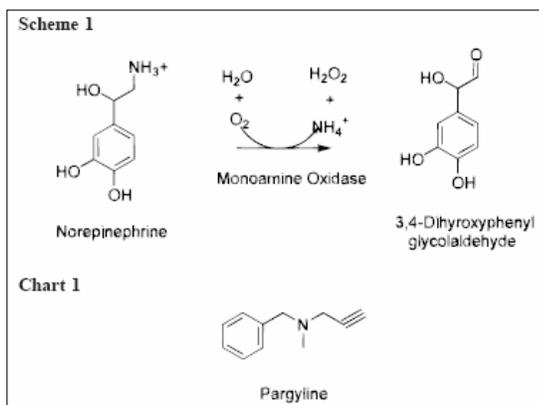
This method, Bradu and Hawkins, finds outliers in a 2-way table. The method is robust to multiple outliers in the table.

Here there are clearly unusual data points. The central dense sweep of data is clearly marching to a different drummer. This figure points the scientists to data points that need careful evaluation.

# MAO Data

## Analysis of a Large Structure/Biological Activity Data Set Using Recursive Partitioning<sup>1</sup>

Andrew Rusinko, III,<sup>†</sup> Mark W. Farnen,<sup>‡</sup> Christophe G. Lambert,<sup>§</sup> Paul L. Brown, and S. Stanley Young<sup>\*</sup>



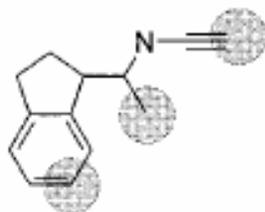
27

A key conceptual idea is that the data set in question might contain compounds that exert their effects through different mechanisms.

Careful biochemical work at Abbott many years ago led to the discovery of two types of MAO inhibitors. A data set containing multiple modes of action is extremely difficult to make sense of using standard statistical methods, e.g. linear regression. The basic problem is that features important for one mechanism are likely unimportant for a second mechanism. Each mechanism dilutes the measured importance of features for the other mechanism.

Recursive partitioning, available in JMP, progressively separates the compounds into groups and within the groups it is then possible to identify important features.

## Active Compounds found from ACD



AGN113

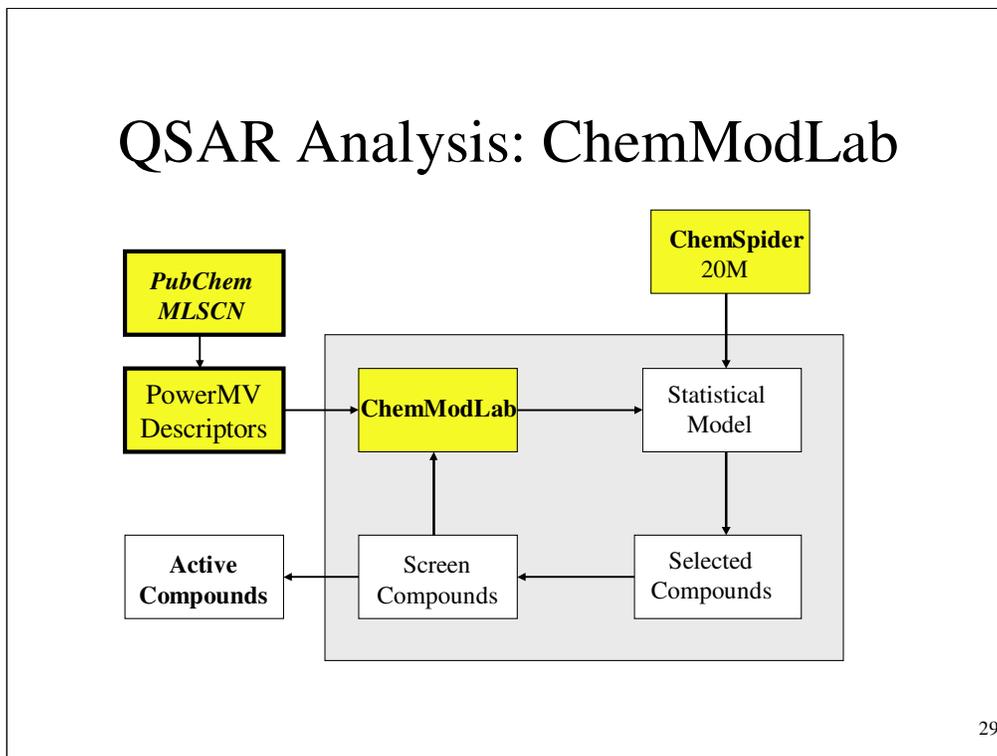


ACPHENEN2

28

Here are compounds from the two mechanisms with important features marked.

# QSAR Analysis: ChemModLab



29

Sequential screening is a standard strategy for lead finding in drug discovery. We make three major contributions.

1. ChemModLab, using descriptors from PowerMV, is freely available and it provides easy to use sophisticated statistical modeling.
2. ChemSpider has a large collection of compounds for virtual screening.
3. ChemSpider offers access to chemical structures from literature papers so that it can be used to construct a training data set.

We use a PubChem data set to demonstrate methods.

## ChemModLab Statistical Methods

### **Methods include:**

- Trees: RandomForest, rpart, tree
- Neural networks
- k-nearest neighbors
- Support vector machines
- Partial least squares
- Partial least squares with linear discriminant analysis
- Least angle regression
- Ridge regression
- Elastic net
- Principal components regression
- *Family ensemble of k-nearest neighbors, using 70% selection*
- *Family ensemble of tree, using 70% selection*
- *Family ensemble of rpart, using 70% selection*
- *randomForest using 70% selection*

30

Here is a list of statistical methods available. R is extensively used.

## Demonstration Example

---

### *Plan*

Use PubChem data set, AID460.

Use PowerMV to compute atom pair descriptors.

RandomForest analysis computed in ChemModLab.

V screen 2M / ~20M e-compounds from ChemSpider.

31

Data, Experiment AID 460, is taken from PubChem. Descriptors, atom pairs, were computed in PowerMV. Our experience is that atom pairs and RandomForest consistently perform well in prediction so these were chosen for modeling activity. We used all 100 active compounds and about 5,000 inactive compounds for the modeling.

## ChemModLab – High Performance Computing

1. ECCR “owns” 12 processors. There are 600 64-b processors. (~1000 total)
2. We use 64-b R.
3. We optimize the job loading for our multiple-model computing.

ChemSpider demonstration trial:

1. 2M compounds put into 42 batches of ~50k each.
2. Each batch takes <3 minutes of processing.
3. 42 batches on 12 processors takes about 12 minutes.

32

A great deal of effort has been put into optimization of this process. We work on algorithms. We also work on distributed processing. We have moved to 64 bit R to allow better use of memory. In a real project, time is money. A day saved is worth ~\$80k. We are still working on how to move to larger training data sets.

# AID460

The Penn Center for Molecular Discovery

Cathepsin L

57,821 compounds tested

100 nominally active compounds

48 / 100 retested active

33

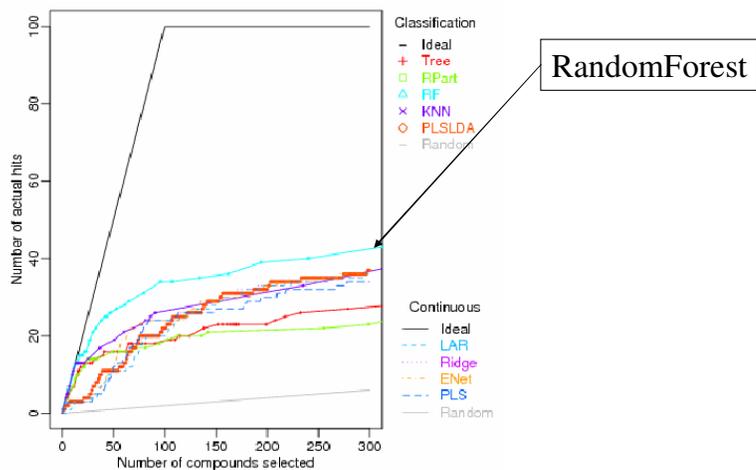
After discussion with the U Penn staff, AID460 was chosen as the training data set.

Note that there were 52 “false positives”. Hopefully our descriptors and modeling method will capture the “signal” in the data set.

Preliminary analysis indicated that atom pairs had predictive power.

## ChemModLab Results

- *Accumulation curves: compare methods, atom pairs 546*



During our research we have tested a number of classification methods and RandomForest often does quite well.

Indeed RandomForest performed well for this data set. RF was able to find ~40 of the ~50 true actives.

## Results

2M compounds were taken from the ~20M compound collection in ChemSpider.

These were placed in batches of 50k compounds each for prediction.

Predictions were consolidated and the top 1000 predicted compounds selected.

Results were inspected using PowerMV.

35

Using a computer farm, ChemModLab was able to virtually screen 2M compounds in batches of 50k in about 12 minutes.

## PowerMV, 1000 predicted actives

The screenshot displays the PowerMV V0.72 software interface. The title bar reads "PowerMV V0.72". The menu bar includes "File", "View", "Analysis", "Tools", "Window", and "Help". The toolbar contains icons for "New", "Open", "Table", "SDF", "List", "Text", "Explorer", and "Home". The Explorer window shows the file "1000\_sdf.NoHydrogen(1000)" and two other files: "AID460 SD Active noH(100)" and "729558(10)".

The main workspace is a grid of 12 chemical structures, numbered 1 through 12. The structures are arranged in a 3x4 grid. The first row contains structures 1, 2, 3, and 4. The second row contains structures 5, 6, 7, and 8. The third row contains structures 9, 10, 11, and 12. Each structure is a chemical structure diagram.

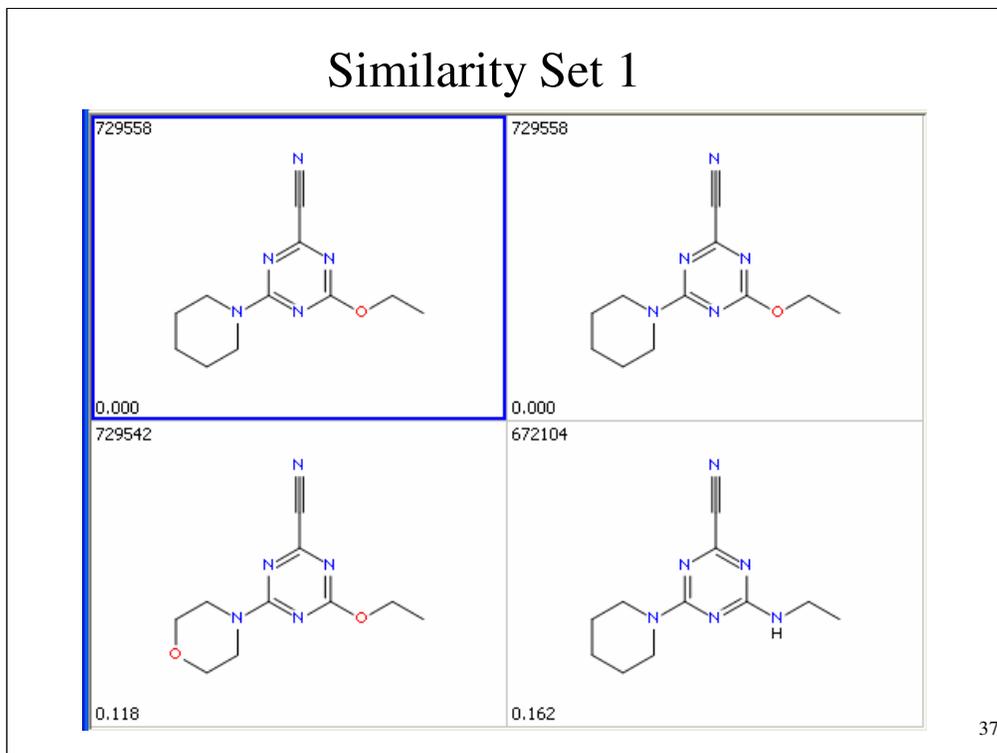
On the left side, there is a "Layout" panel with the following settings:

- Layout: Task | x
- Cell Size: 128, 128
- Column: 4
- Display: (none)
- Rotate: 0
- Search: [ ]
- In: (name)
- Find Next
- Find All
- High Quality
- Show Attributes
- Show Hydrogen
- Show Title

The number 36 is located in the bottom right corner of the screenshot.

PowerMV displays the first 12 compounds in the list.

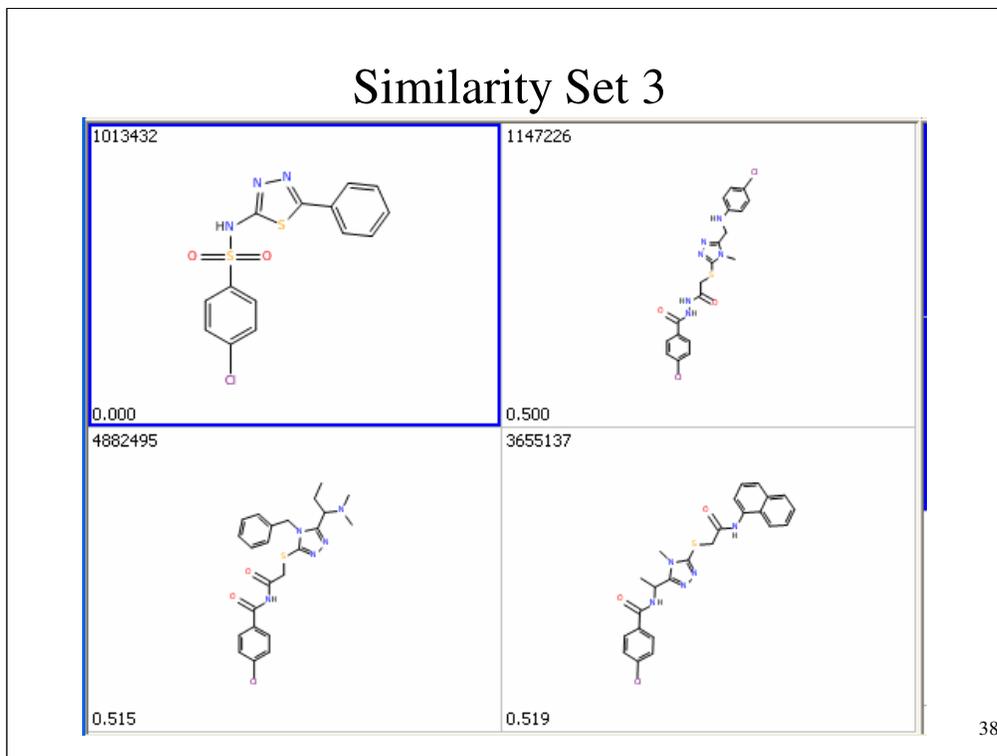
## Similarity Set 1



37

PowerMV can be used for “near neighbor” selection. Here we have one of the active compounds, 729558, and the nearest three near neighbors from the 1000 predicted actives. Note that one of the virtual screening compounds was actually in the ChemSpider collection of 2M compounds. It is a nice check that the similarity searching of PowerMV can actually find itself. Also note that one can flip and rotate compounds to put them into the same orientation.

## Similarity Set 3



38

Here are the three nearest neighbors from the top 1000 to compound 1013432, one of the actives in the training set. All four compounds have the para-chloro pheny ring. All have a NN containing five membered ring. Etc. If the e selected compounds prove active, these could be considered “lead hopping”.

## Software and Web Services

For JMP SMILES, contact [young@niss.org](mailto:young@niss.org).

PowerMV at [www.niss.org/PowerMV](http://www.niss.org/PowerMV)

Google (NCSU ChemModLab) to get to  
a very sophisticated QSAR analysis system.

Questions ?