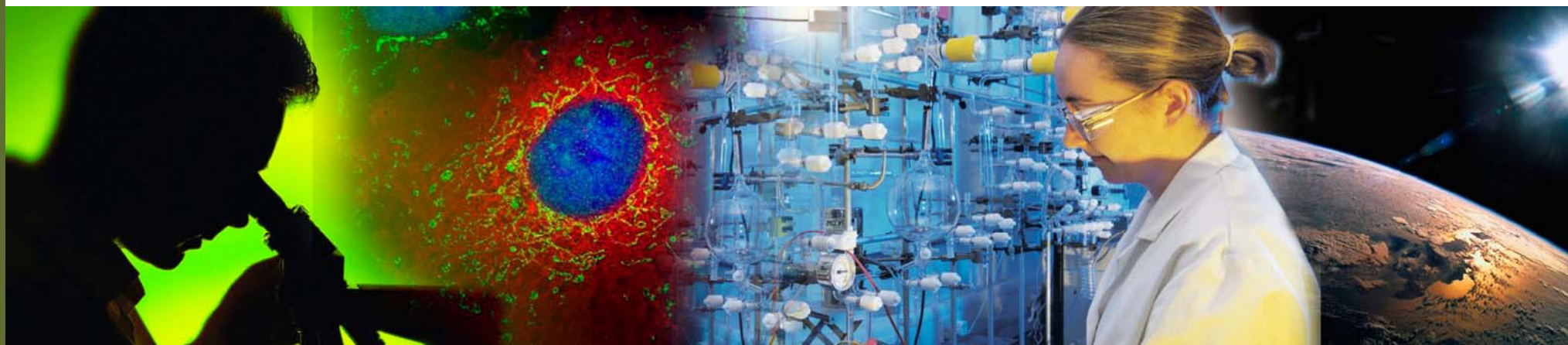




Many Patterns & Many Methods

New methods for visualising & utilising multiple analysis techniques in polymorph and salt screening systems

²⁰⁰⁹
Gordon Barr, Chris Gilmore & Gordon Cunningham
WestCHEM, Chemistry Department, University of Glasgow



High throughput screening experiments can generate hundreds of PXRD patterns a day

Problems with:

Data quality.

Sample quality.

Data quantity.

Need for automation, and speed.

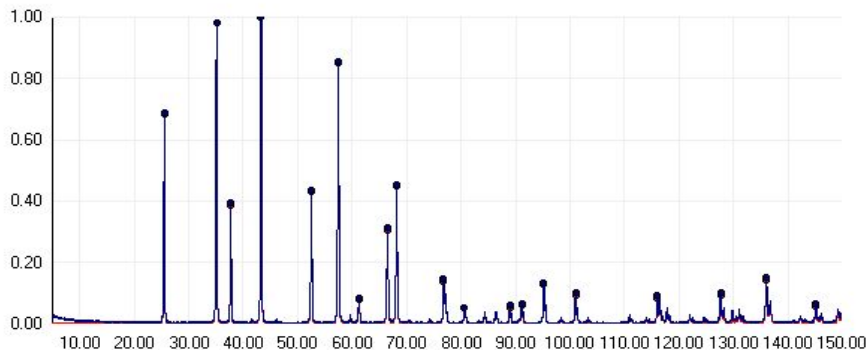
How do you deal with hundreds of samples from a single technique (e.g. XRPD), let alone more than one at once?

Compare pairs of patterns using full-profile parametric and non-parametric statistics

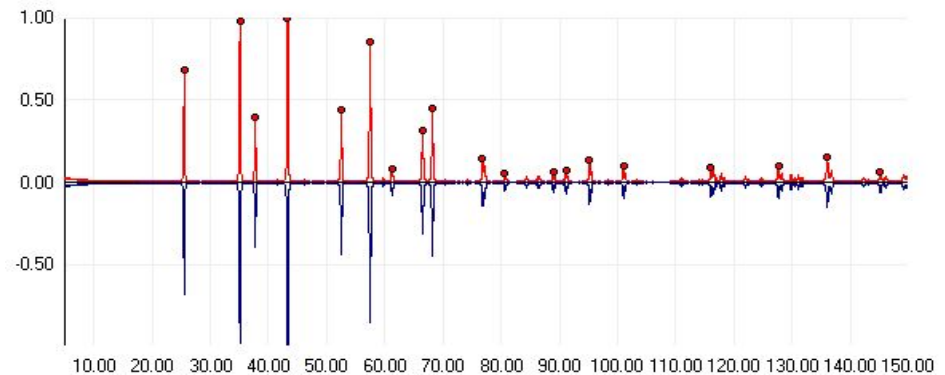
Match every data point – forget about the peaks!

Use correlation coefficients:

- Pearson correlation coefficient (parametric).
- Spearman correlation coefficient (non-parametric).

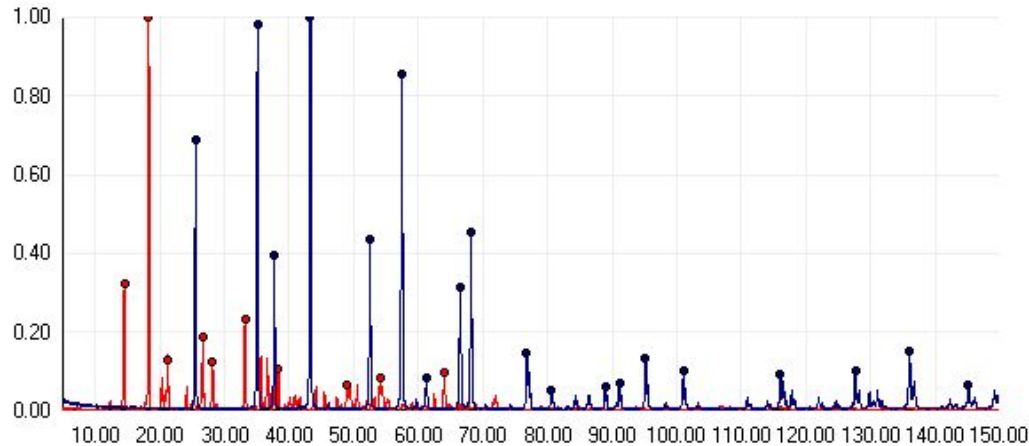


Correlation coefficient $\approx +1.0$



Correlation coefficient ≈ -1.0

Match two patterns:



-> Get a correlation coefficient

Pattern A matches Pattern B with a correlation of:
0.314

Match n patterns:

	upd3-004.c	upd3-005.c	upd3-006.c	upd3-007.c	upd3-008.c	upd3-010.c	upd3-014.c	upd3-015.c	upd3-016.c	upd3-017.c
upd3-	1	-1.958017E	-1.340543E	-1.788726E	-1.675198E	-6.313147E	-1.085791E	-1.579619E	5.26832E-C	6.930218E
upd3-	-1.958017E	1	0.856992	-2.006913E	0.925139	-0.0159797	-2.706579E	-5.597485E	7.042118E	0.0161234
upd3-	-1.340543E	0.856992	1	-1.078085E	0.940968	-1.085994E	-1.904628E	-7.160438E	-6.845203E	-5.209728E
upd3-	-1.788726E	-2.006913E	-1.078085E	1	6.384568E	-1.180575E	-2.069083E	-9.256598E	-6.849316E	-6.128258E
upd3-	-1.675198E	0.925139	0.940968	6.384568E	1	-1.396583E	-2.449129E	-8.501349E	-8.519993E	-6.499929E
upd3-	-6.313147E	-0.0159797	-1.085994E	-1.180575E	-1.396583E	1	0.733269	3.687835E	3.371778E	0.2762358
upd3-	-1.085791E	-2.706579E	-1.904628E	-2.069083E	-2.449129E	0.733269	1	0.1159216	8.542597E	0.548274
upd3-	-1.579619E	-5.597485E	-7.160438E	-9.256598E	-8.501349E	3.687835E	0.1159216	1	0.1743946	0.2046806
upd3-	5.26832E-C	7.042118E	-6.845203E	-6.849316E	-8.519993E	3.371778E	8.542597E	0.1743946	1	0.548263
upd3-	6.930218E	0.0161234	-5.209728E	-6.128258E	-6.499929E	0.2762358	0.548274	0.2046806	0.548263	1

-> Get a correlation between every pair of patterns

-> can build a $n \times n$ correlation matrix

Have a correlation matrix

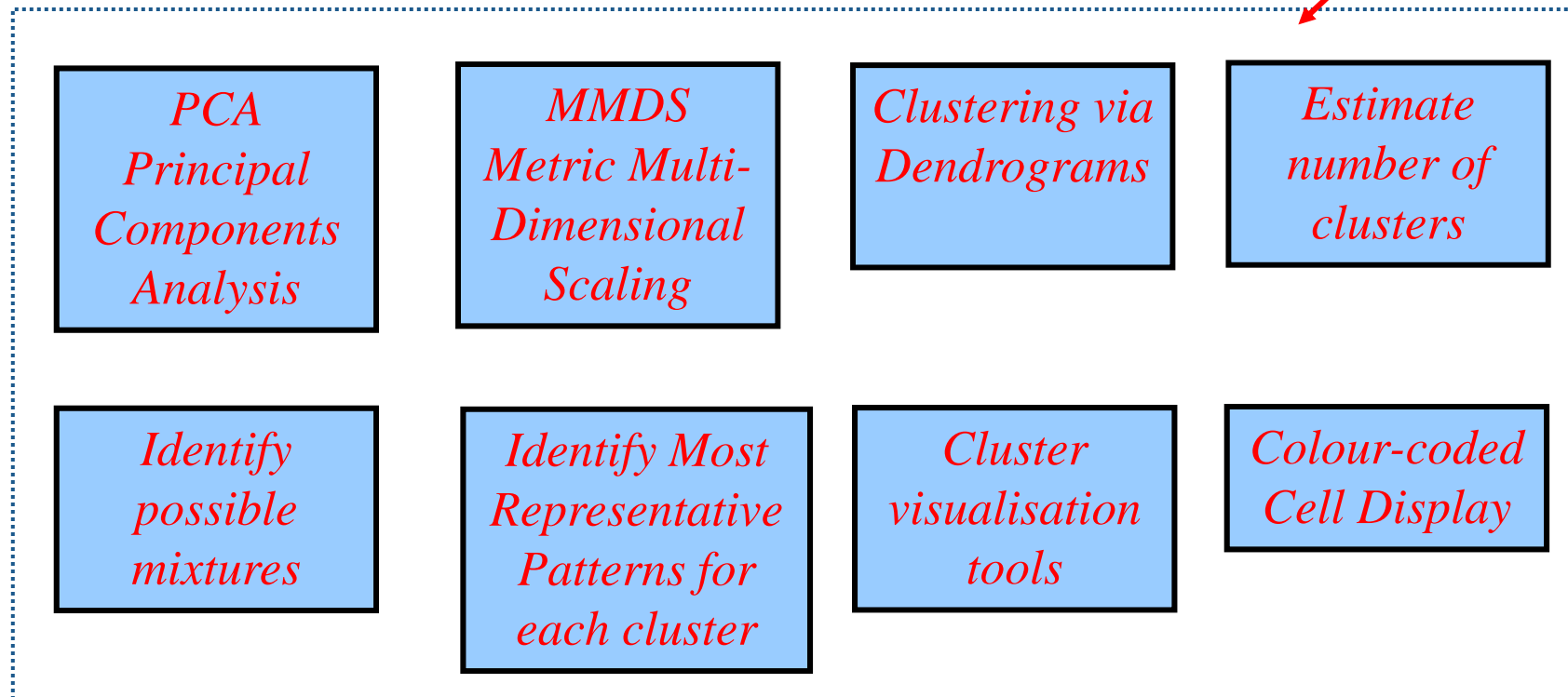
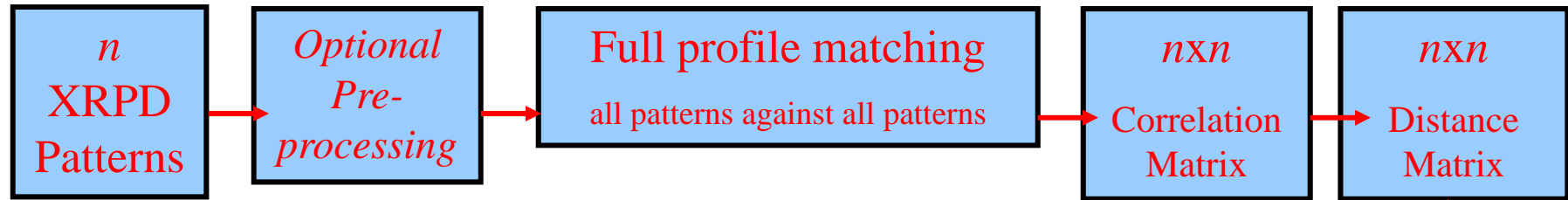
Convert correlations to distances:

- Correlation = 1.0 \Rightarrow distance = 0.0
- Correlation = -1.0 \Rightarrow distance = 1.0
- Correlation = 0.0 \Rightarrow distance = 0.5

Take the distance matrix and perform:

Cluster analysis, Principal components analysis, Metric multidimensional scaling, Fuzzy clustering, Minimum spanning trees *etc.*

To find '*interesting*' patterns and to visualize the data.

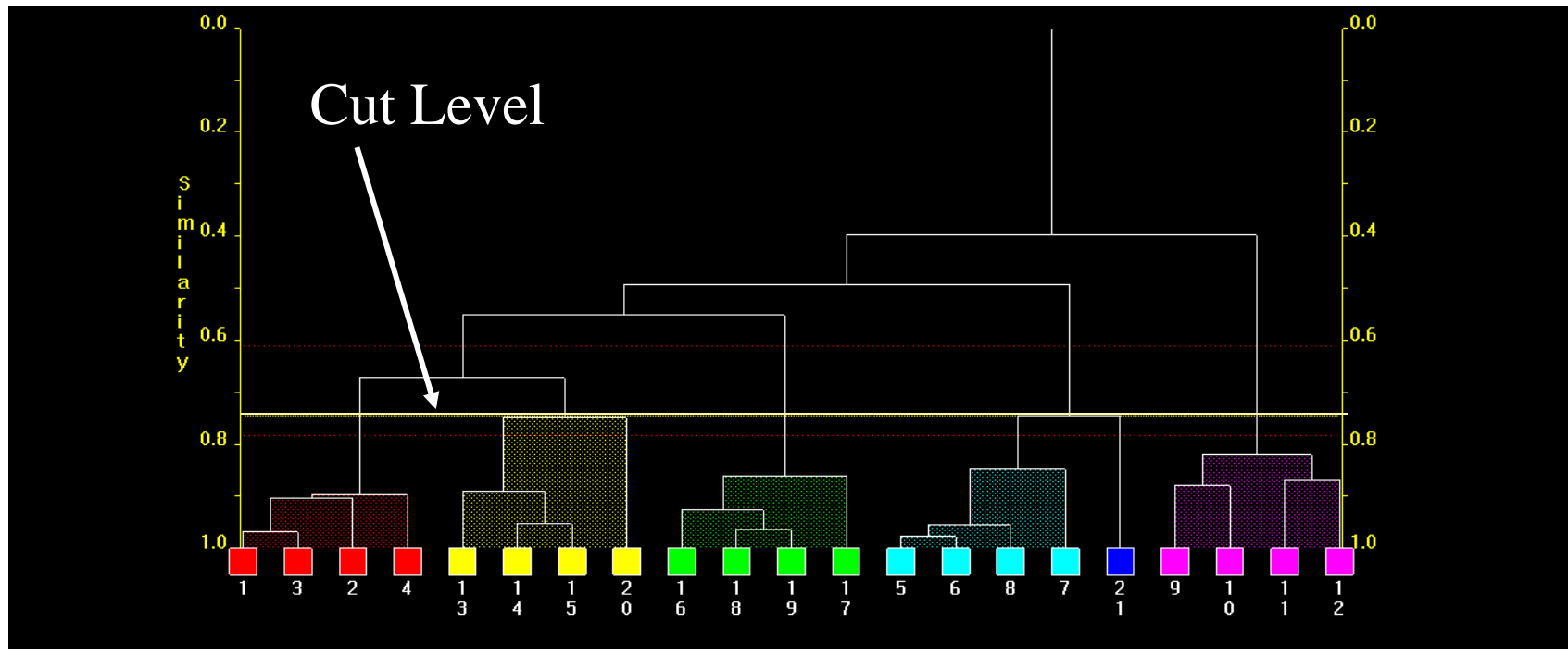


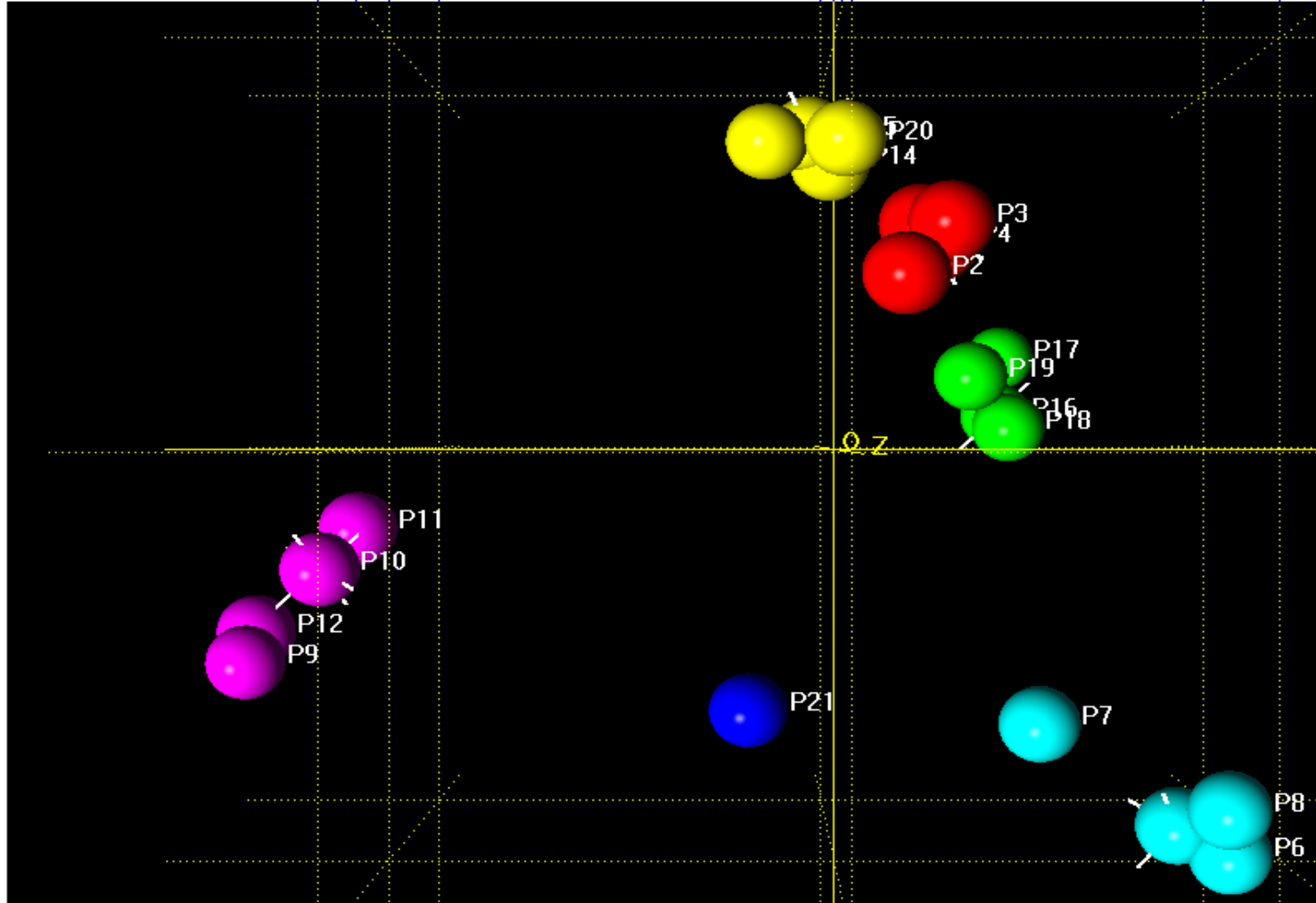
Also indexed as: Cardura XL® , Cardura®

Doxazosin is a member of the alpha blocker family of drugs used to lower blood pressure in people with hypertension.

Doxazosin is also used to treat symptoms of benign prostatic hyperplasia (BPH).

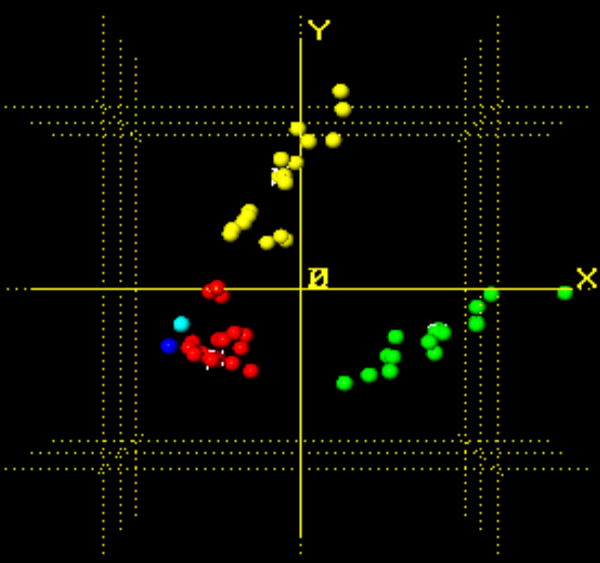
**Study performed using 21 patterns of
5 polymorphic forms of Doxazosin**



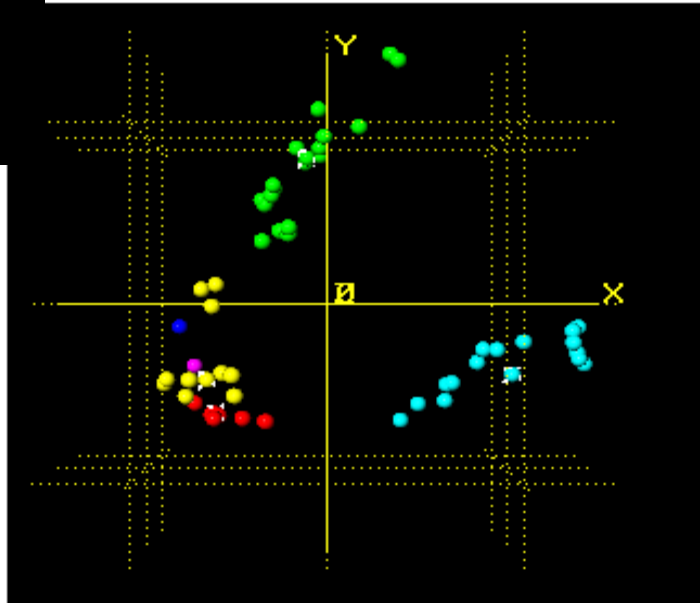


Example: Carbamazepine

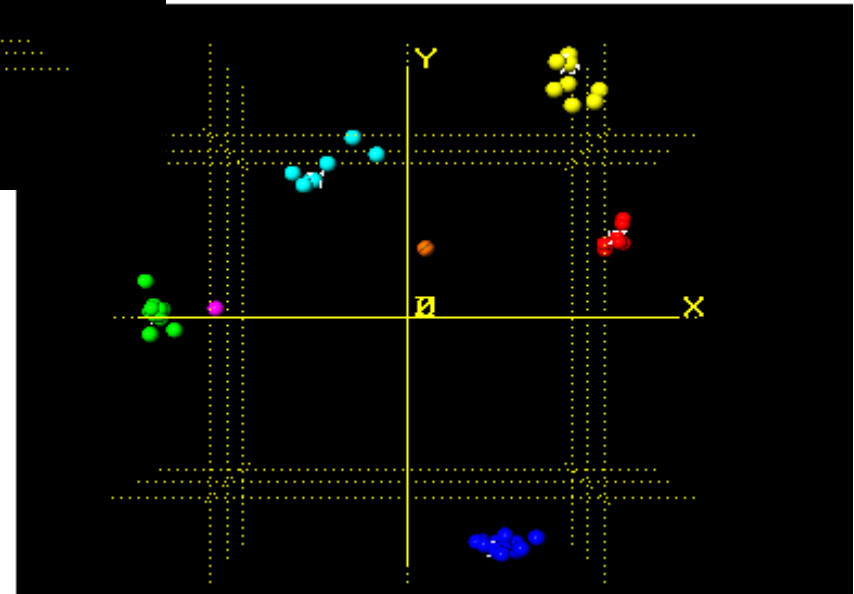
<- No background subtraction

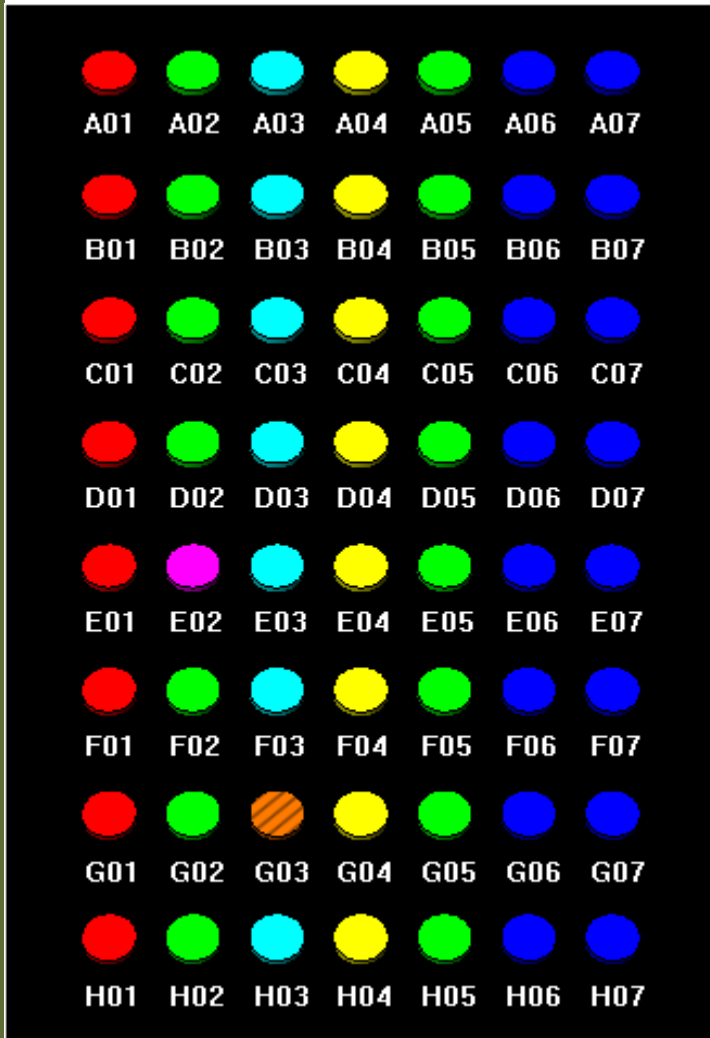


<- Normal background subtraction



Strict background subtraction ->





A1 – H1

Alpha

A2 – H2

Beta

A3 – H3

30 % alpha, 70 % beta

A4 – H4

Gamma-Beta mixture

A5 – H5

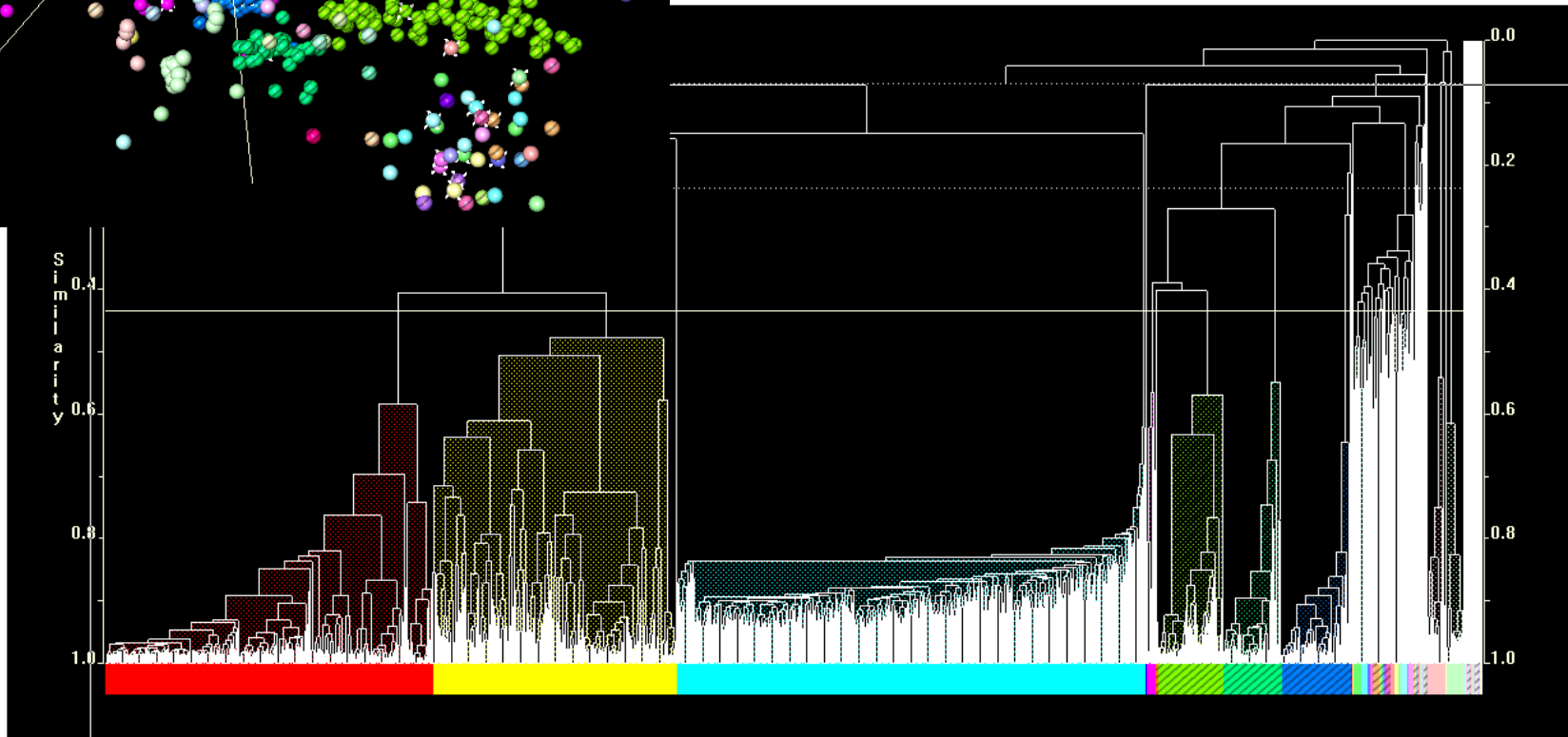
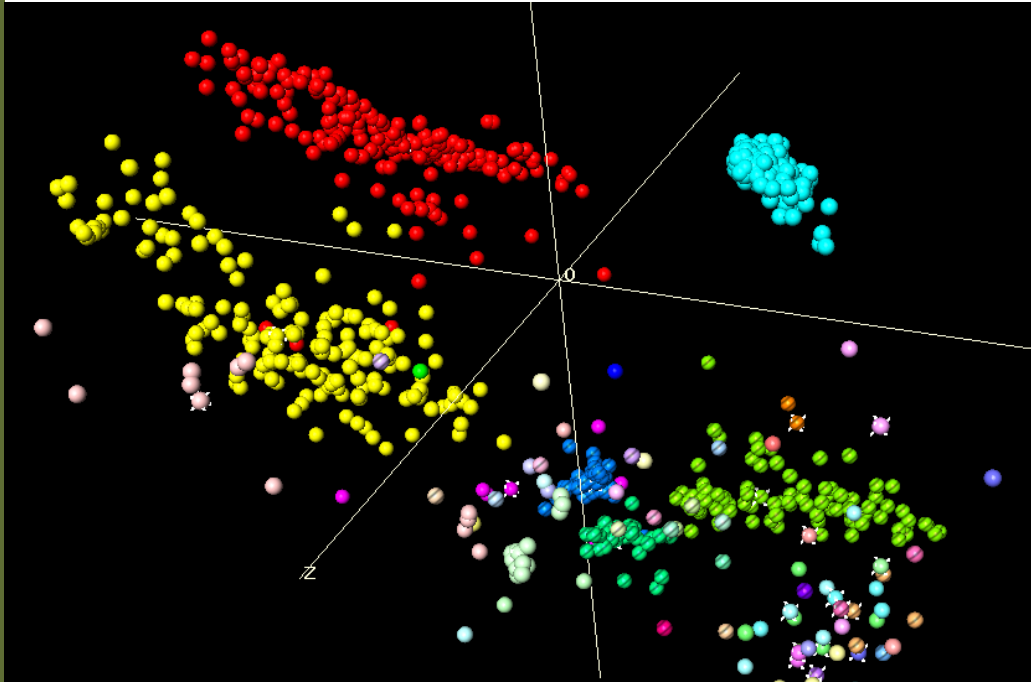
DMSO

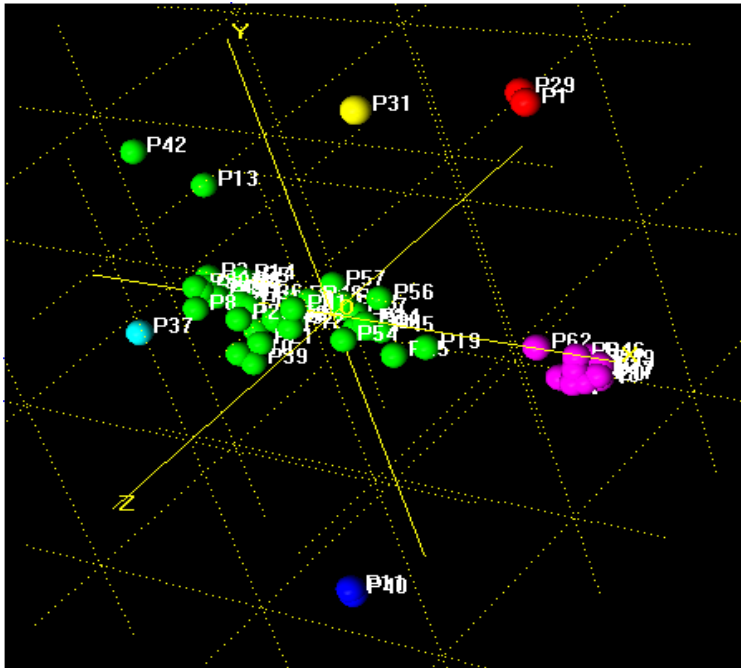
A6 – H6

Dioxane 1-4 Form II

A7 – H7

Dioxane



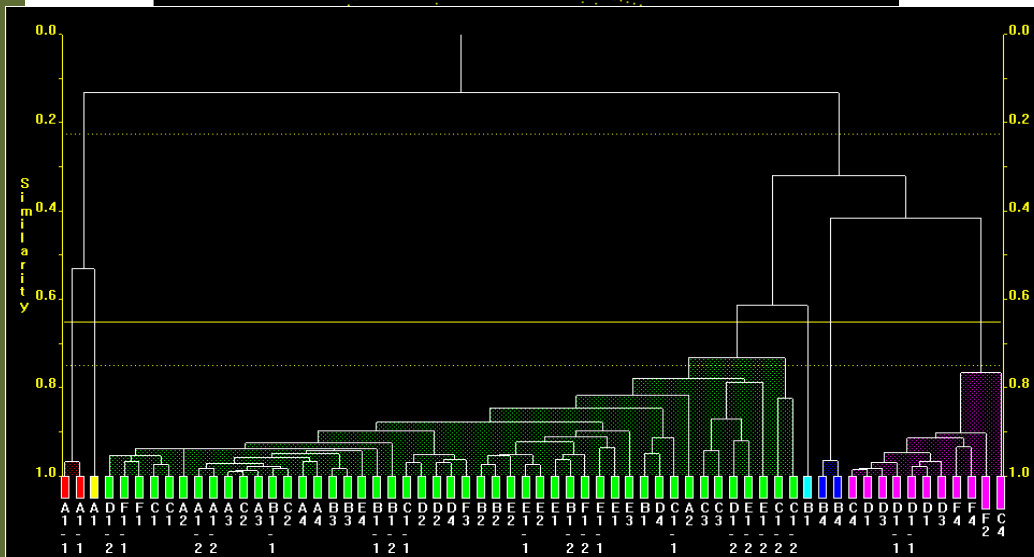


... I'd cast my eye over the spectra and have done a spectral comparison of the data by eye.

I INDEPENDENTLY came up with five different spectral groups. So bottom line is PolySNAP using background subtraction routines gave EXACTLY the same result as me doing a spectral comparison by eye.

....thought you all should know that IMHO this is a significant step forward.

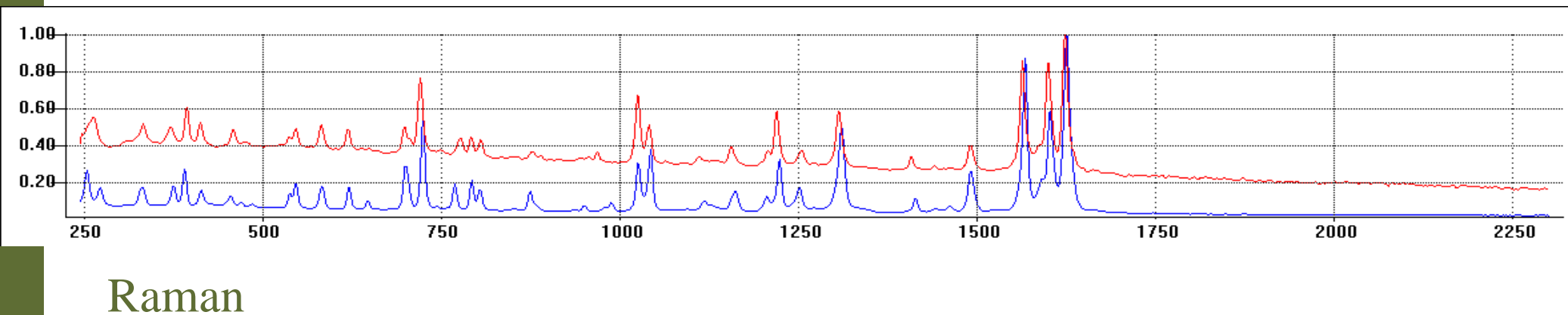
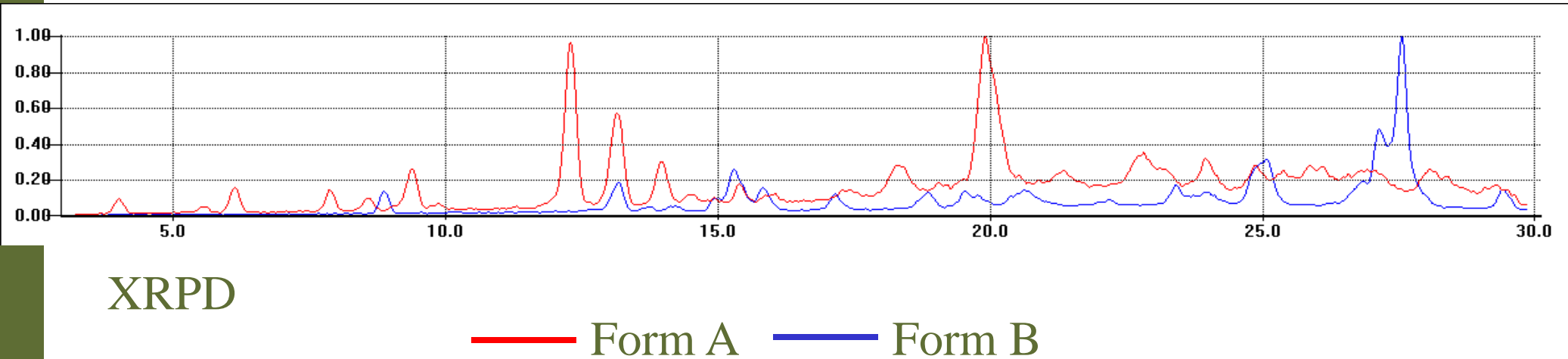
Don Clark, Pfizer Global R&D

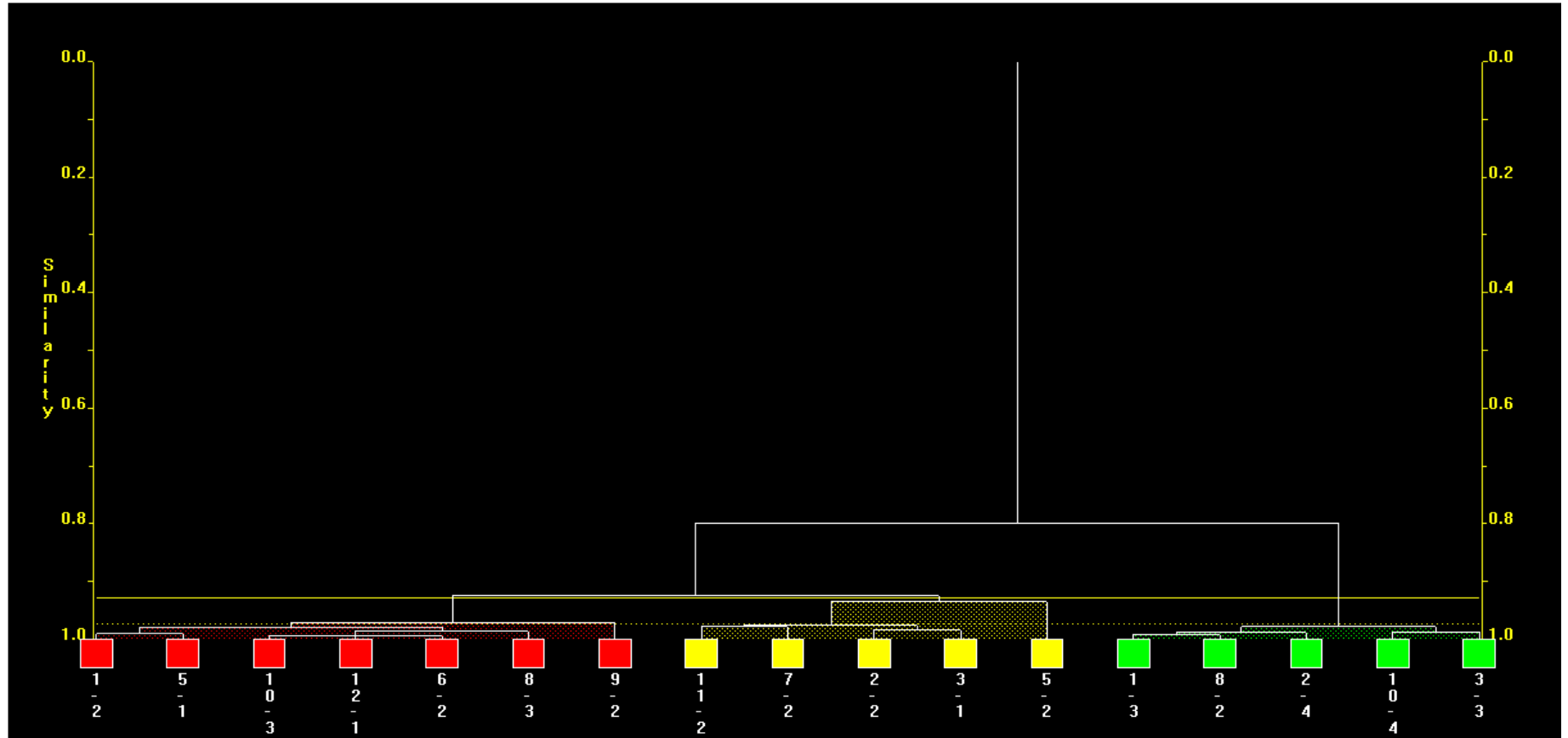


Different background types

Much smaller differences between patterns

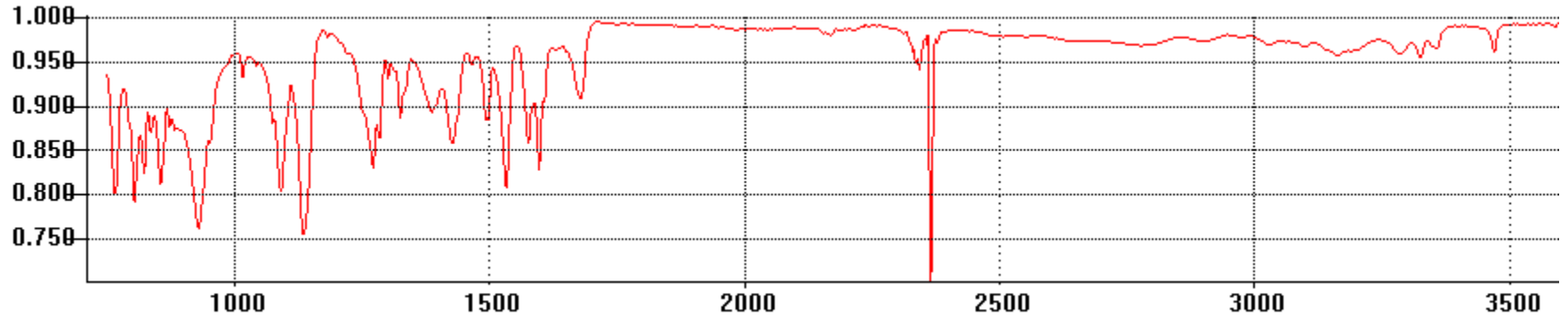
Cosmic spike problems



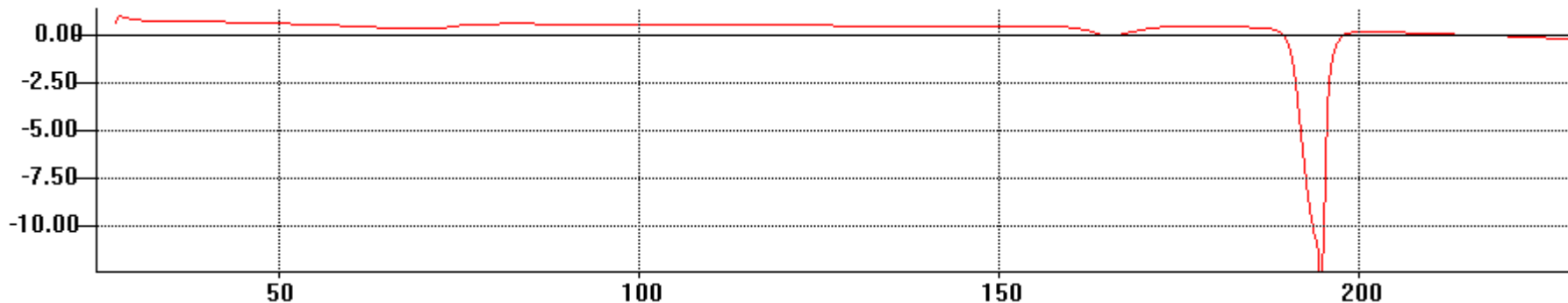


Doesn't have to be PXRD or Raman data:

IR



DSC



Other Profile Data

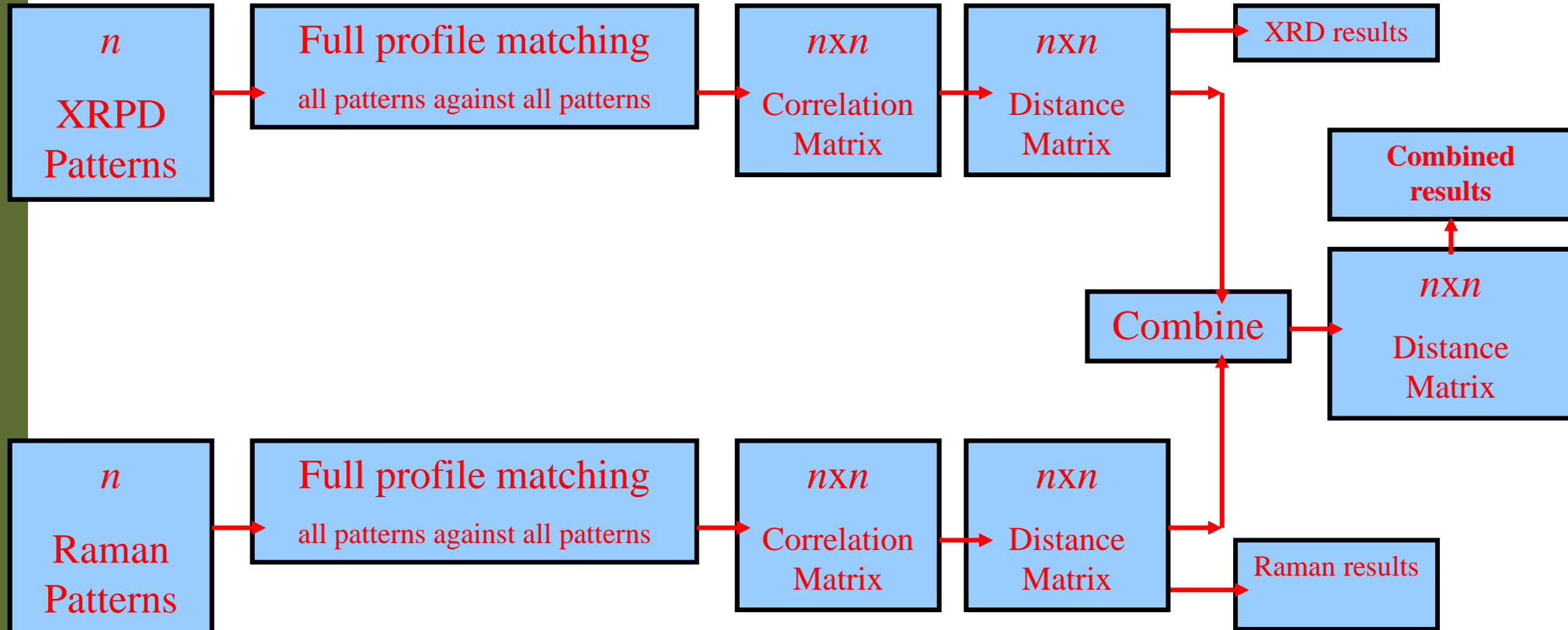
Numeric Data

XRF

Combined XRPD + Raman instruments now available

Applying multiple techniques to the same samples helps give additional information to work with

How would we actually combine results from two (or more) such different techniques ?



Manual weighting:

Give a single weight to each dataset as a whole

Combine datasets on that basis

- e.g. Powder 0.8, Raman 0.2

Dynamic weighting:

Automatically calculate optimal weighting for each entry in each dataset

Each data set has a 2-D distance matrix d

D_k is square ($n \times n$) distance matrix for dataset k

e.g. we have Raman and XRPD data on 20 samples, so $k = 2$, $n=20$.

We want a Group Average Matrix G to optimally describe our data

Specify diagonal weight matrices W_k which can vary over the k datasets

We want to minimise
$$\sum_{k=1}^K \left\| \mathbf{B}_k - \mathbf{G} \mathbf{W}_k^2 \mathbf{G}' \right\| \quad (1)$$

Where
$$\mathbf{B}_k = -\frac{1}{2} (\mathbf{I} - \mathbf{N}) \mathbf{D}_k (\mathbf{I} - \mathbf{N})$$

(a double-centering operation on D), and $\mathbf{N} = \mathbf{1}\mathbf{1}' / n$

Solve (1) to get best values for G and W

Bryne, D.V., O'Sullivan, Dijksterhuis, G.B., Bredie, W.L.P. & Martens, M. (2001) Food Quality Pref. 12, 171-187.

Attempt to develop a sensory vocabulary to describe the flavours of warmed up meat patties (!)

The data were assessed by

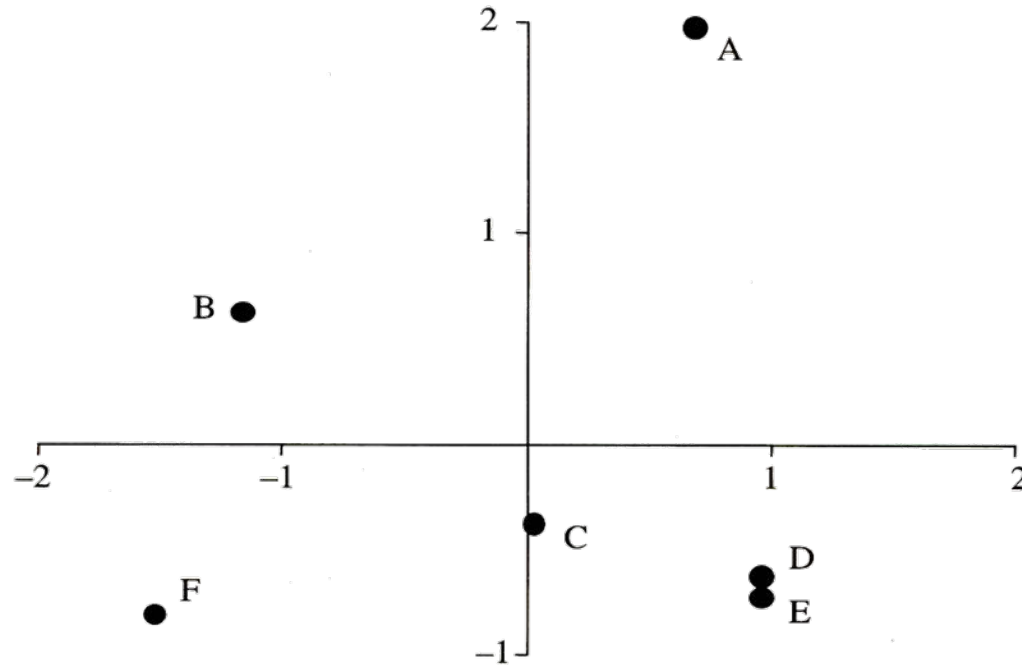
8 assessors (= no. of data types)

judging 6 products (= no. of data sets for each type)

using 20 attributes (= no. of points in each data set).

So you have 8 D matrices. The group average G in 2 dimensions gives:

A, B...F are the six patties. It can be seen that D and E are judged to be very similar but there is a big difference in the others.



Dataset of Sulphathiazol, Carbamazepine + Mixtures

16 samples each had data from:

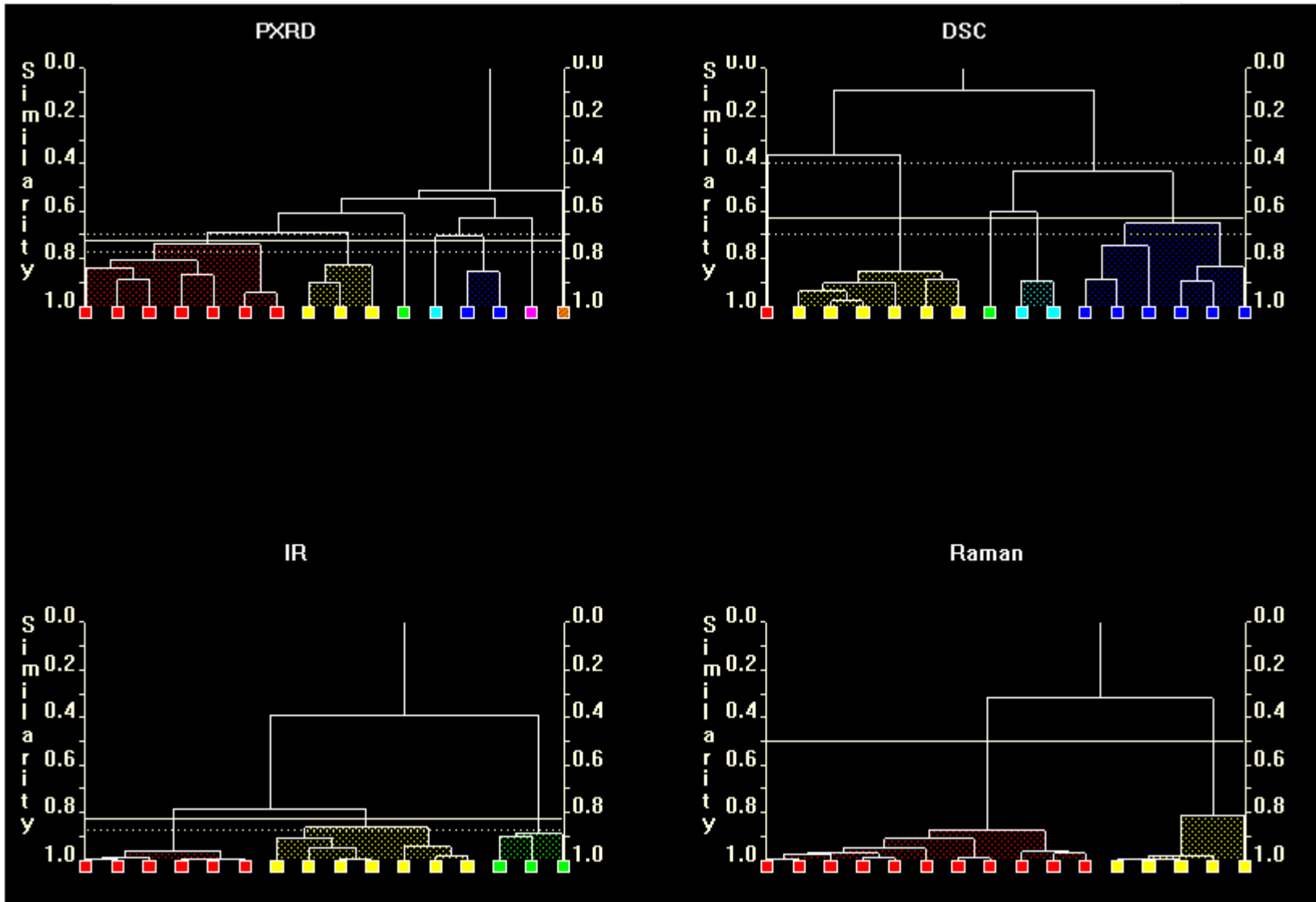
1. PXRD (collected on a Bruker C2 GADDS)
2. DSC (collected on a TA instruments Q100)
3. IR (collected on a JASCO FT/IR 4100)
4. Raman (collected on a Renishaw inVia Reflex)

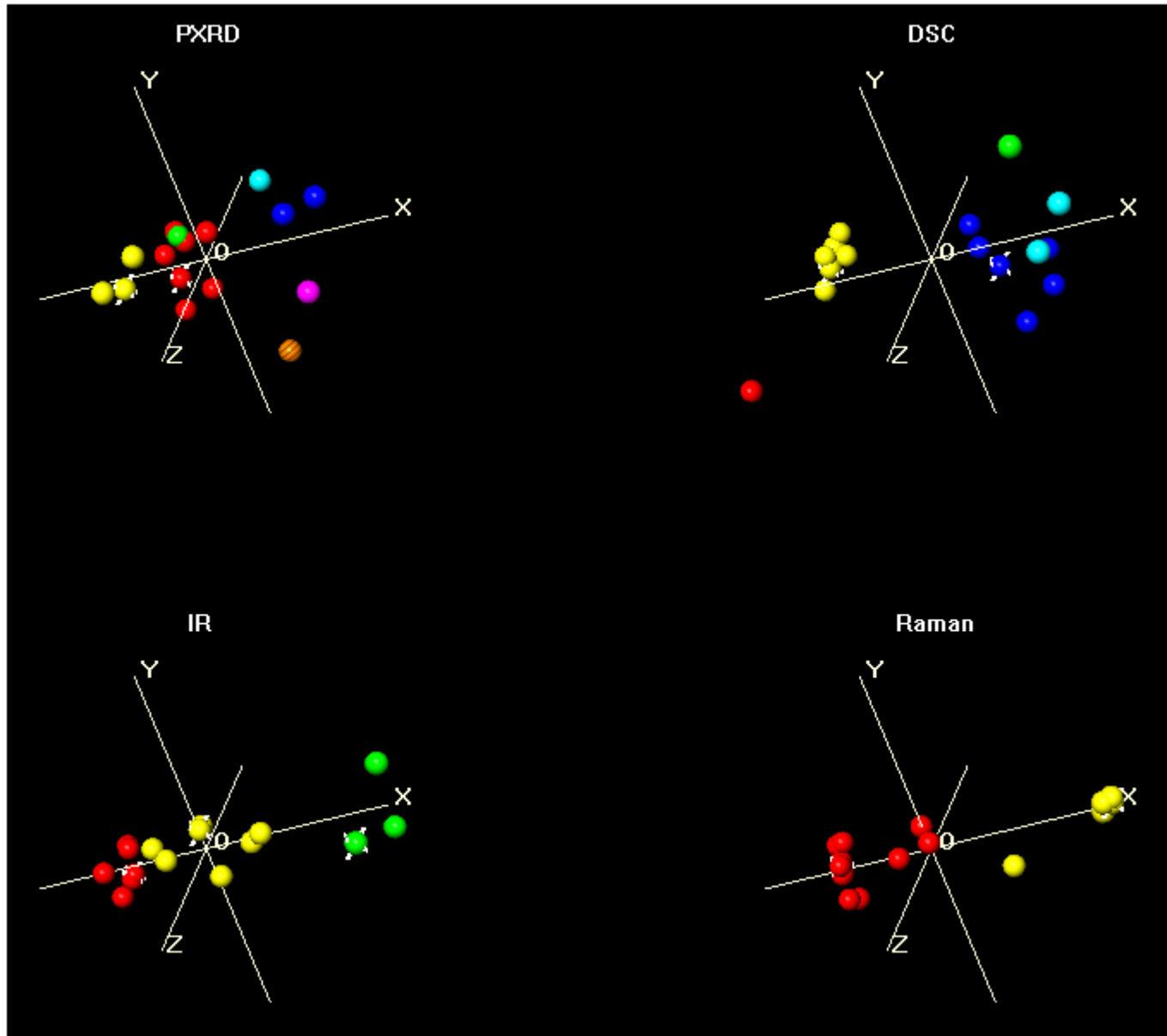
Combinations:

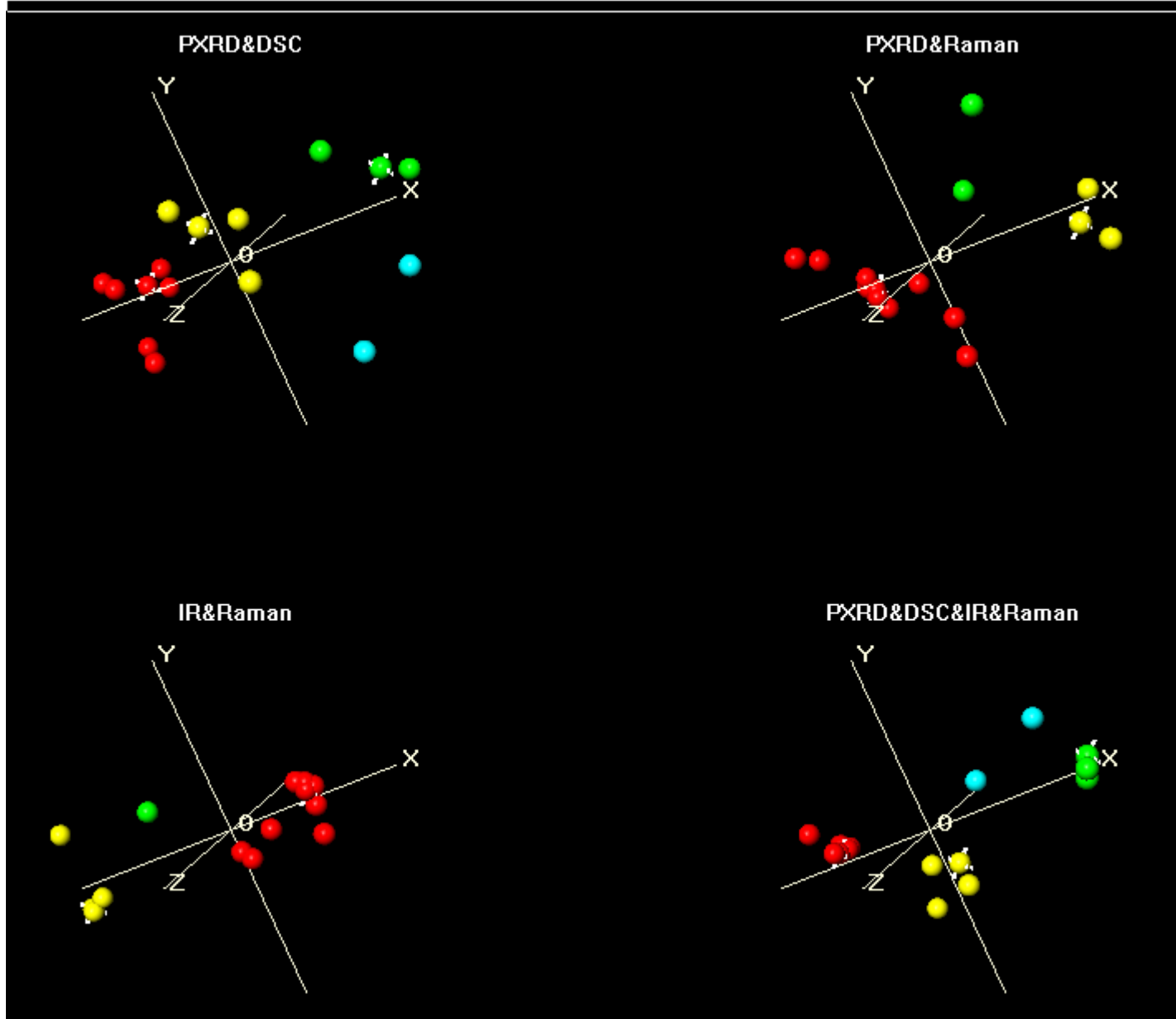
PXRD+Raman

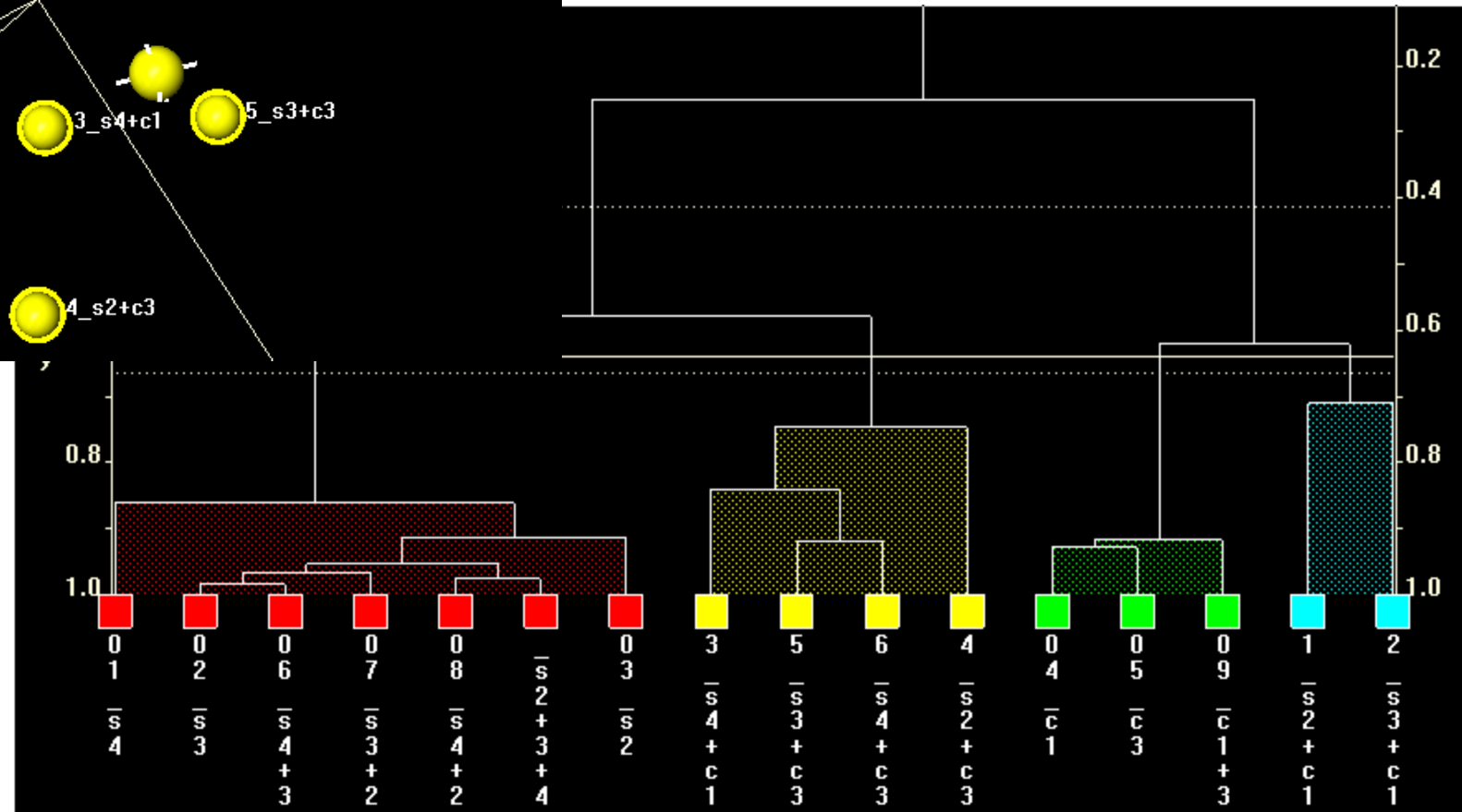
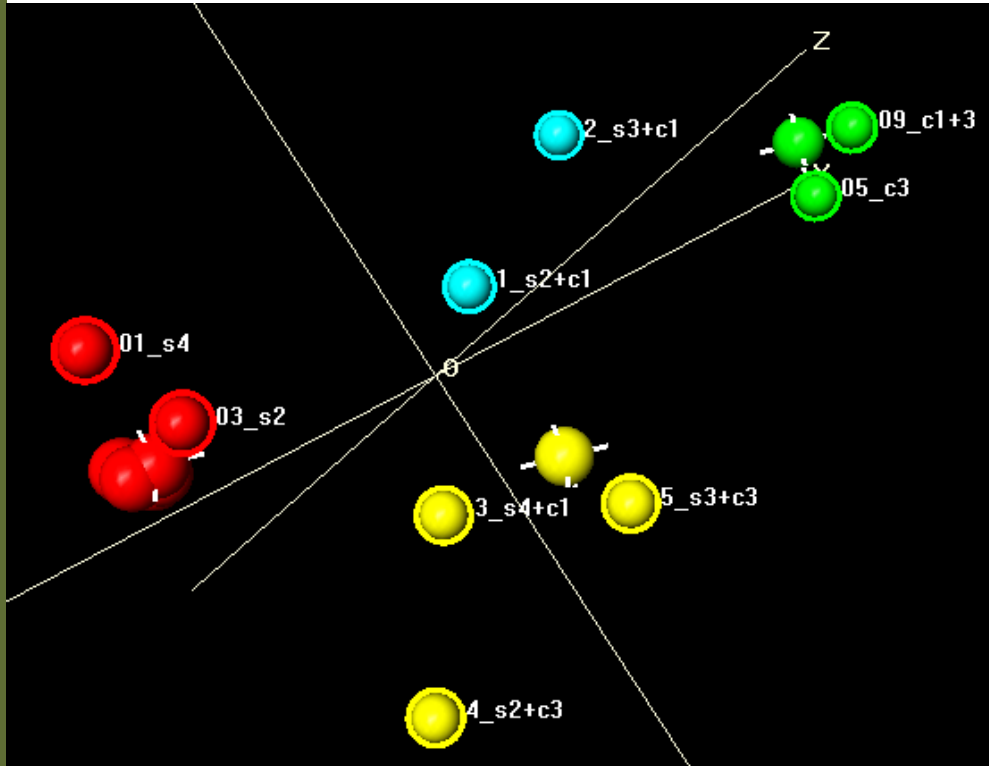
PXRD+Raman+DSC

PXRD+Raman+DSC+IR *etc.* [up to 15 sets of results!]











Full Profile Matching + Cluster analysis methods do very well in distinguishing forms automatically using either Raman or PXRD data individually

Combined results using Dynamic Weighting seem to do better than either PXRD or Raman individually

Use of combined data helps highlight any inconsistencies in separate analyses

- Such inconsistencies would not be obvious with only one data source
- Outliers can then be examined manually in detail

Seeing similar clustering from multiple original data sources increases confidence in the overall results

Full analysis as shown limited to up to 2,000 patterns per data set.

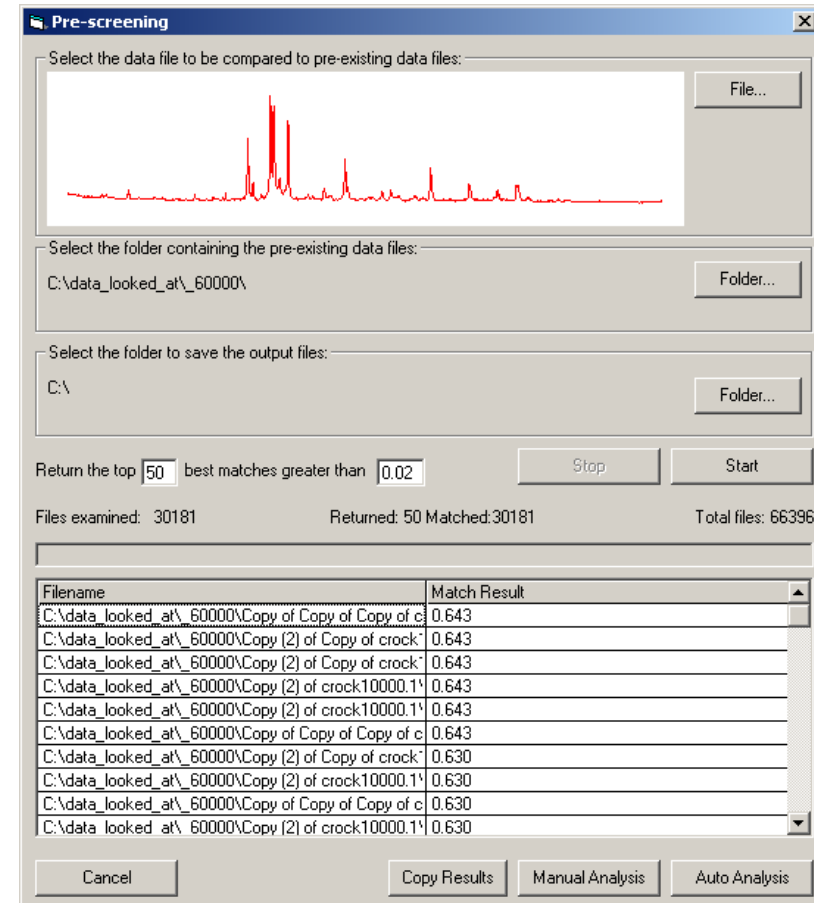
What if you've got more?

Is this new sample something seen before, or new ?

Pre-screening allows a single sample pattern to be compared to large in-house database of existing patterns.

Compare e.g. >66,000 samples to new unknown in ~20 mins

Return the best 50 matches, then visualise using dendrograms, 3D Plots etc as before

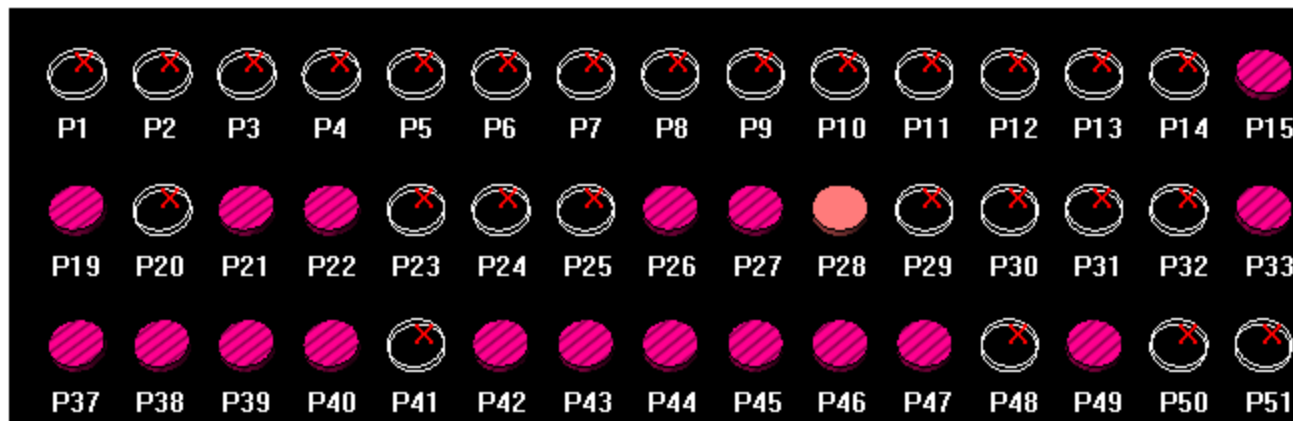


Filename	Match Result
C:\data_looked_at\ 60000\Cop of Copy of Copy of c	0.643
C:\data_looked_at\ 60000\Cop (2) of Copy of crock	0.643
C:\data_looked_at\ 60000\Cop (2) of Copy of crock	0.643
C:\data_looked_at\ 60000\Cop (2) of crock10000.1\	0.643
C:\data_looked_at\ 60000\Cop (2) of crock10000.1\	0.643
C:\data_looked_at\ 60000\Cop of Copy of Copy of c	0.643
C:\data_looked_at\ 60000\Cop (2) of Copy of crock	0.630
C:\data_looked_at\ 60000\Cop (2) of crock10000.1\	0.630
C:\data_looked_at\ 60000\Cop of Copy of Copy of c	0.630
C:\data_looked_at\ 60000\Cop (2) of crock10000.1\	0.630

Salt Screening: not interested in samples consisting of

- One of our starting materials
- Mixture of multiple starting materials
- Given a library of starting materials to compare the new samples to:

Just highlight what's new and interesting



PolySNAP

Matlab or other stats packages

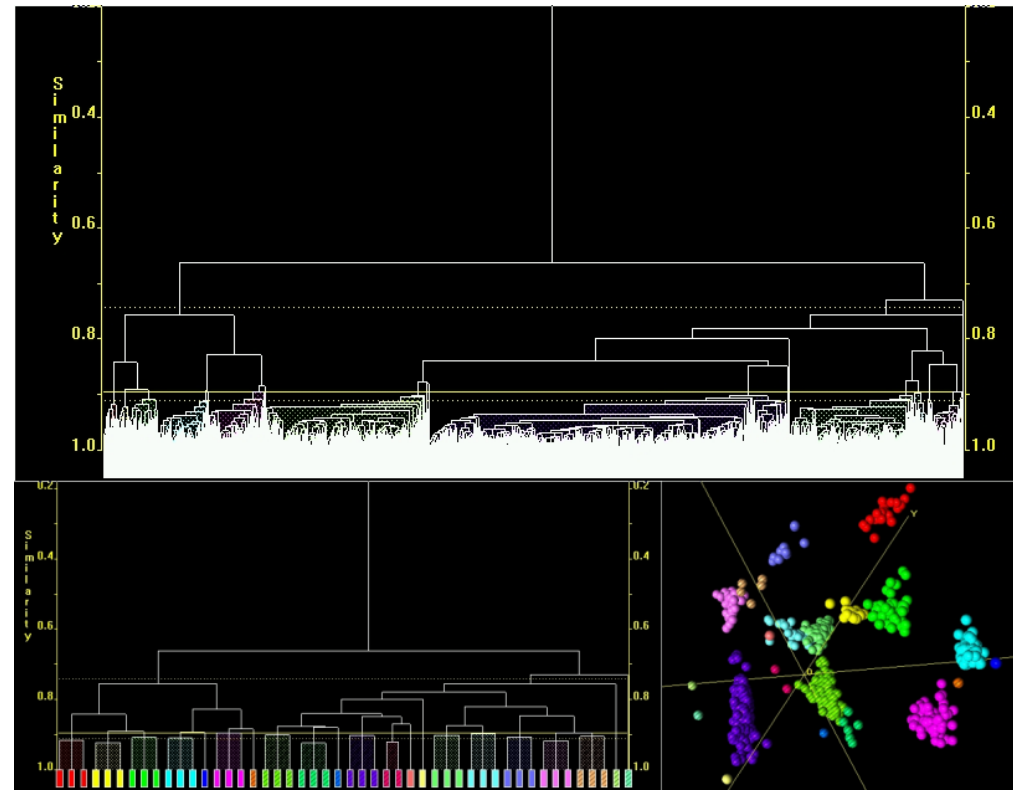
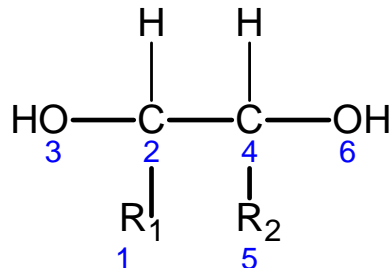
dSNAP

-Cluster & visualise 3D

fragment geometry similarities

from the

Cambridge Structural Database



Many thanks to....

Arnt Kern & Karsten Knorr, Bruker AXS

Chris Frampton & Susie Buttar, Pharmorphix

For more information, please contact us:

Email:

snap@chem.gla.ac.uk

Web:

www.chem.gla.ac.uk/snap