Applying data mining approaches to further understanding chemical effects on biological systems

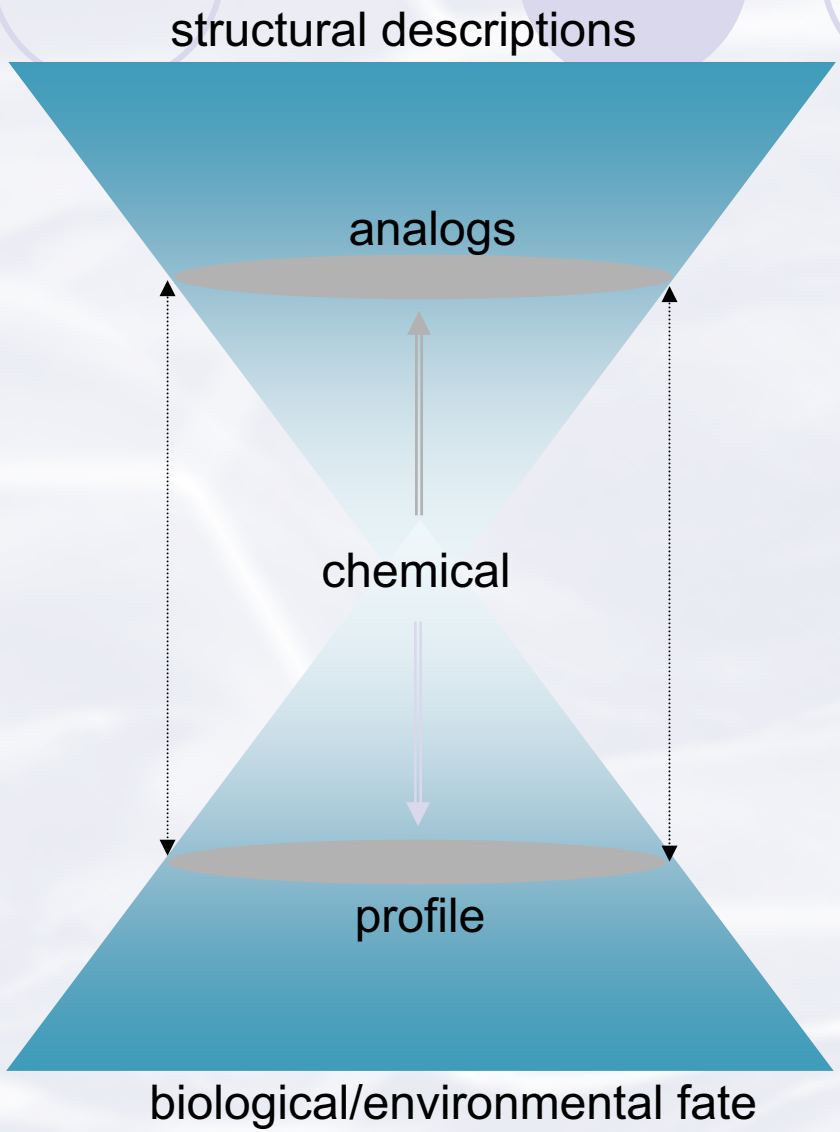
Chihae Yang, Ann Richard, Jennifer Fostel

March 28, 2007

Leadscope®

This paper does not reflect EPA policy.

Linking biology to chemistry



Current Computer-Aided Drug Design, 2006, *2,* 1-19.

NTP High Throughput Screening

- A method to prioritize the selection of chemicals for NTP bioassays
  - Test chemical nomination procedure
    - rank or rule out “most likely toxic” and “most likely not toxic” compounds
- An approach to potentially obtain insights for mode of actions
- Predictive toxicology?

NTP HTS – Chemicals and assays

- Structures 1408 chemicals – 1340 selected in this study
- Biological assays
  - Activity: NCGC 10 strains
    - Caspases - 3T3, BJ, SHSY5, H4IIe ,Hek293, HepG2, HUVEC, Jurkat, N2A
    - IkB signaling, JNK Alpha
  - Viability: Fred (7 strains), NCGC (13 strains)
    - 3T3, BJ, SHSY5, H4IIe ,Hek293, HepG2, HUVEC, Jurkat, N2A
  - “Other assays” – NCGC (4 assays)
    - SKNSH, MRC5, Renal, Mesenchymal

Toxicity data used in this study

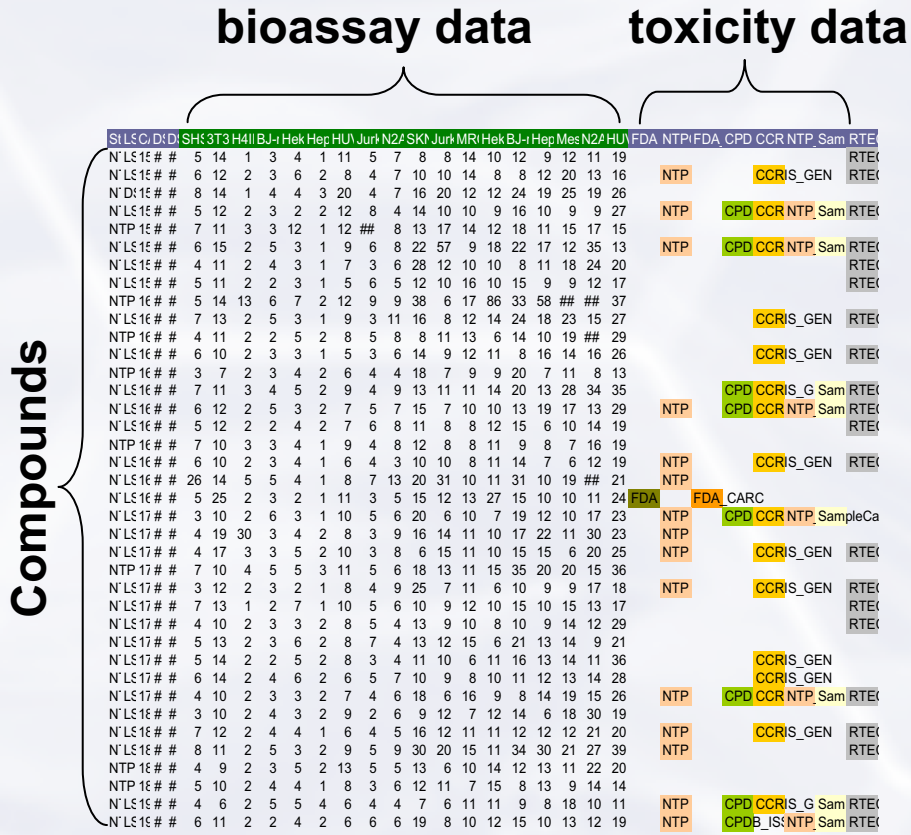
- Carcinogenicity
  - NTP rodent bioassays – mm, fm, mr, fr
  - FDA – mm, fm, mr, fr
  - ISSCAN – mouse, rat
- Genetic toxicity
  - Salmonella strains
  - Mammalian mutagenesis – mouse lymphoma, HPRT
  - In vitro chromosome aberration
  - In vivo micronucleus…
- Rodent acute toxicity
  - RTECS
- Endocrine receptor assay
  - NTCR

Filtering reactive groups

In 1340 chemicals in this study

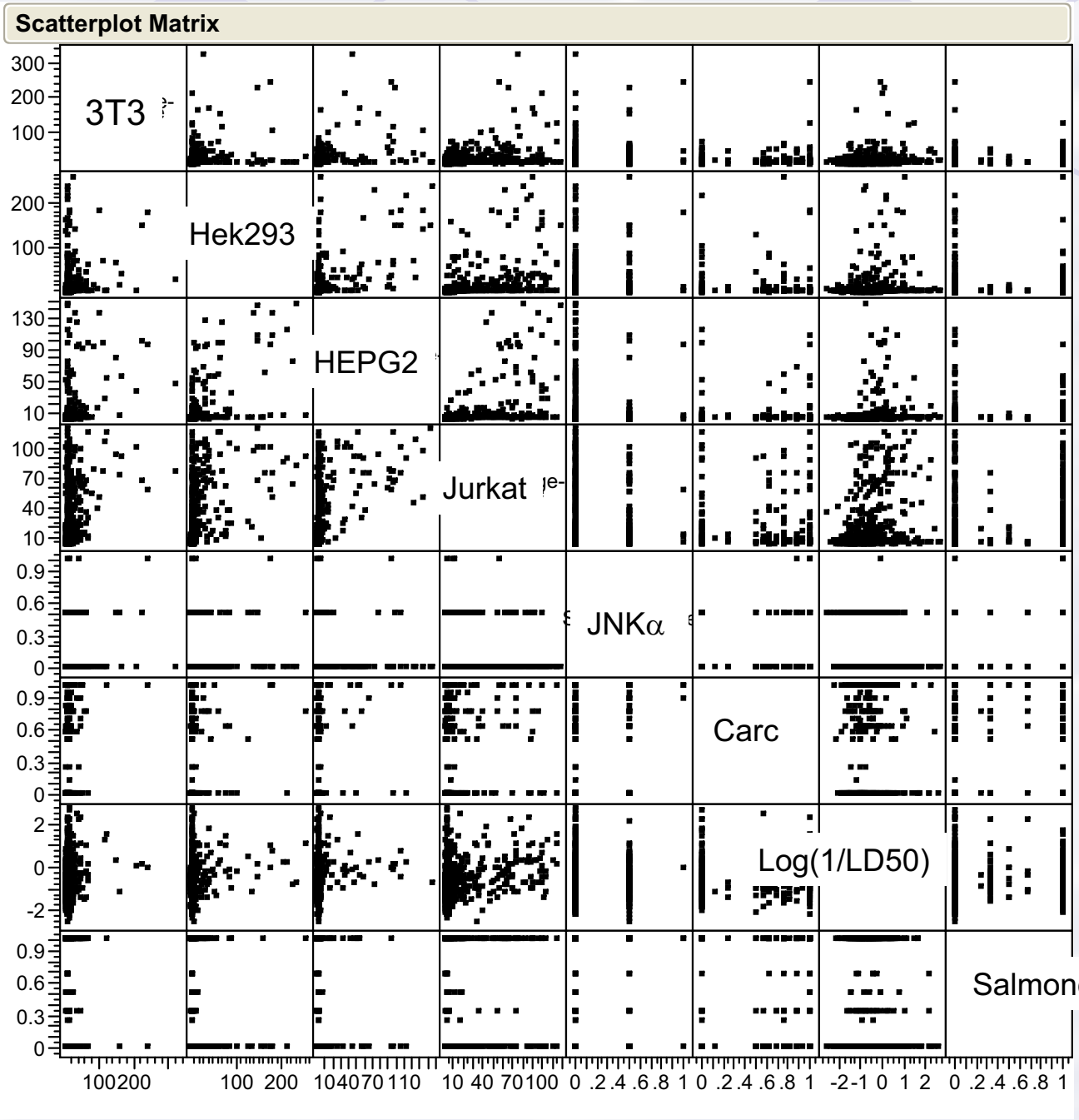
Protein reactions	
Aldehydes	42
α,β unsaturated carbonyl (Michael acceptors)	74
α,β diketones	4
Dinitrohalobenzenes	6
Alkylating agents	
epoxide	33
sulfonate esters	4
sulfonyl halides	2
nitrogen mustard	4
acyl halides	14

Sparse toxicity data space



Toxicity endpoints	counts
LD50 rodent	905
Carcinogenicity call	318
in vivo micronucleus	45
in vitro chromosome aberration	401
Mammalian mutagenesis	314
Salmonella mutagenesis	942

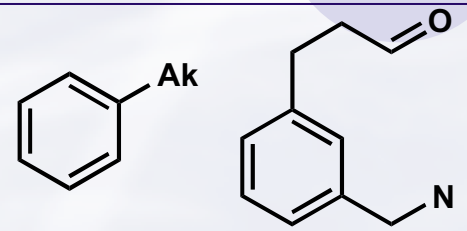
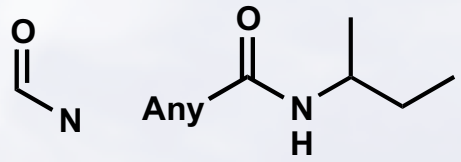
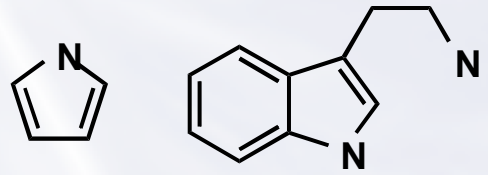
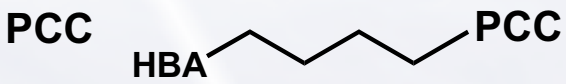
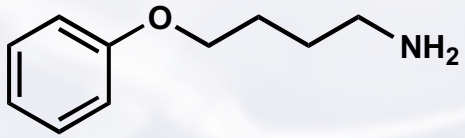
at compound level – not enough tox study overlaps.



multivariate correlation

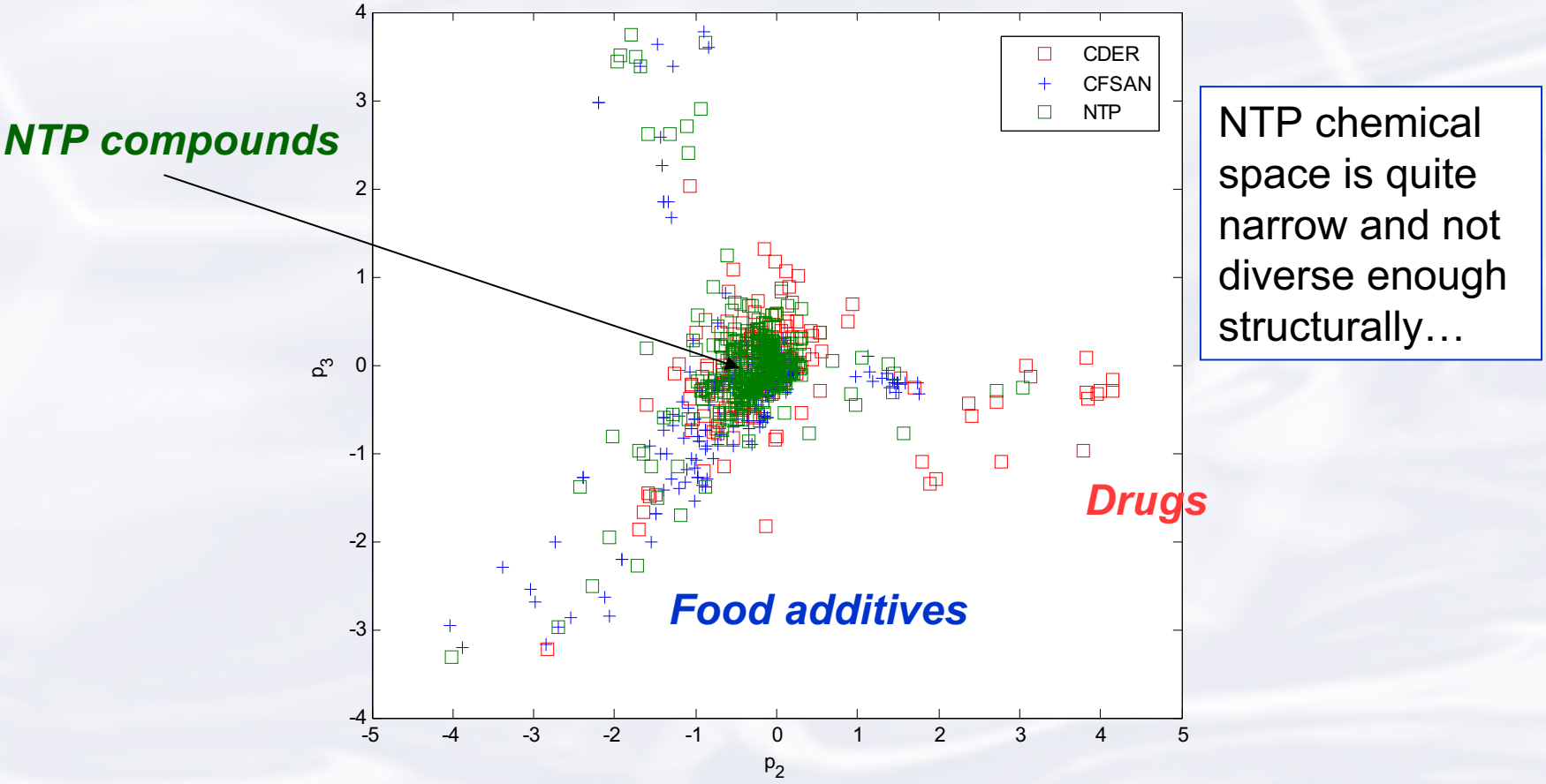
At compound level, virtually not possible to correlate any activities or toxicity data.

Representing structures with Leadscope molecular descriptors

Benzenes	
Functional groups	
Heterocycles	
Pharmacophores	
Spacers	
User defined features	knowledge addition

Applying a set of molecular descriptors to expand chemistry space.

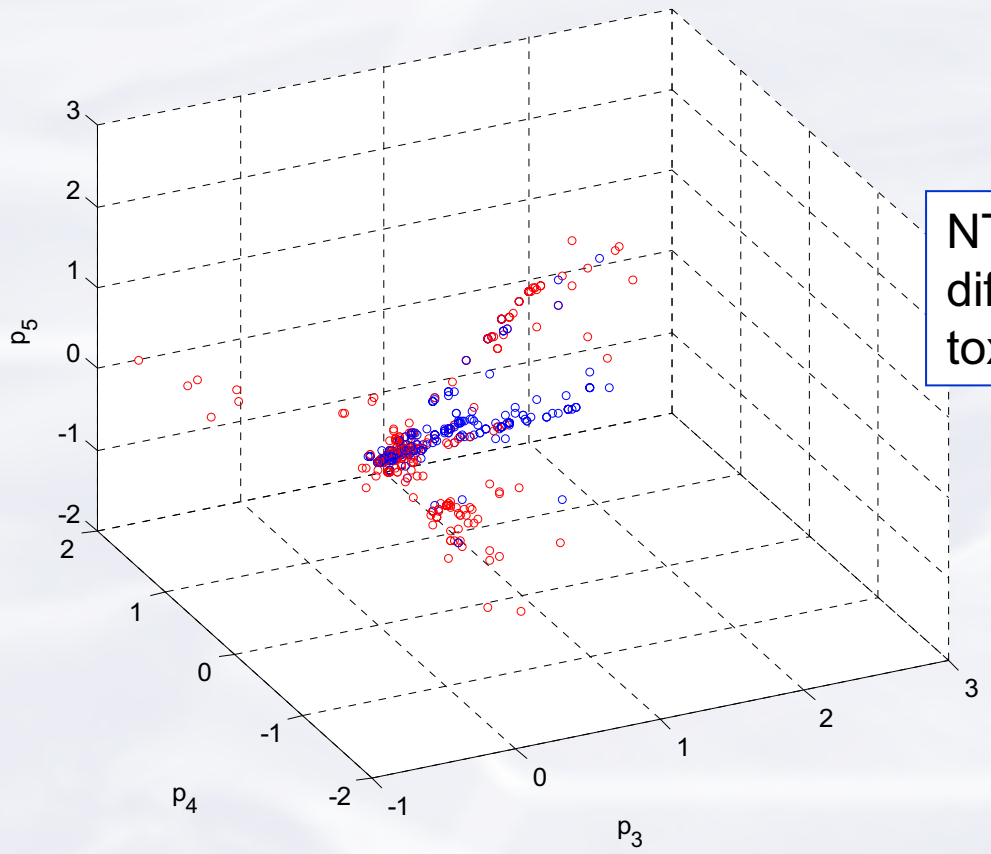
Structure domain characterization by principal component analysis



NTP chemical space is quite narrow and not diverse enough structurally…

using selected features present in >10 compounds

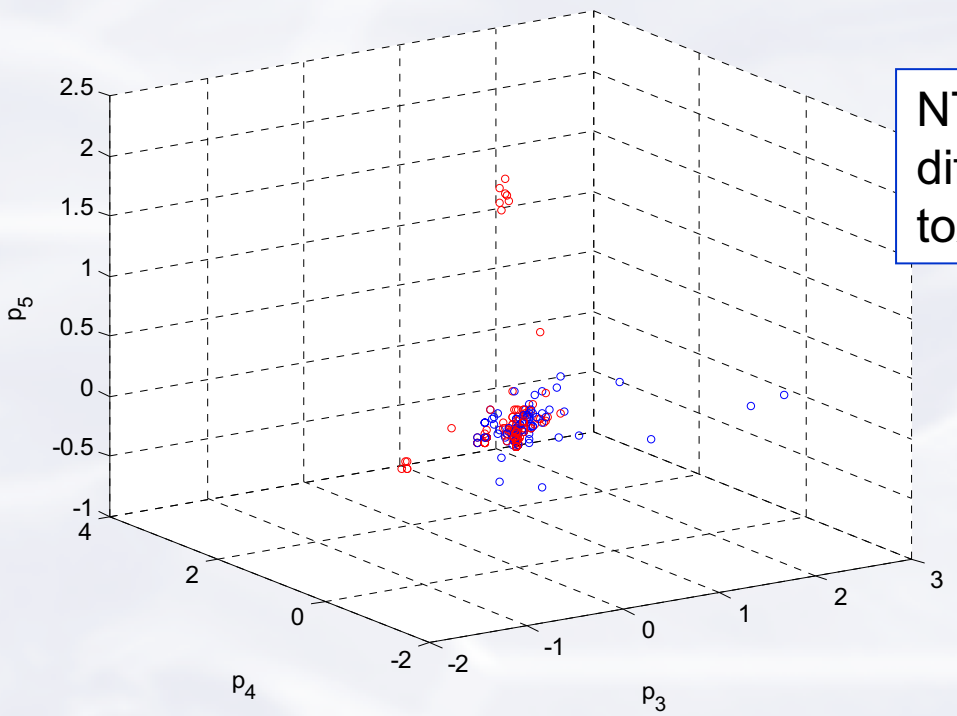
Principal Components – discrimination of NTP chemical space by mutagenic potential



NTP space can be differentiated by toxicity endpoints.

Principal component scores: features (F=100 features selected by chi-sq)
red=Salmonella positive (p>0.5, blue=Salmonella negative (p<=0.5)

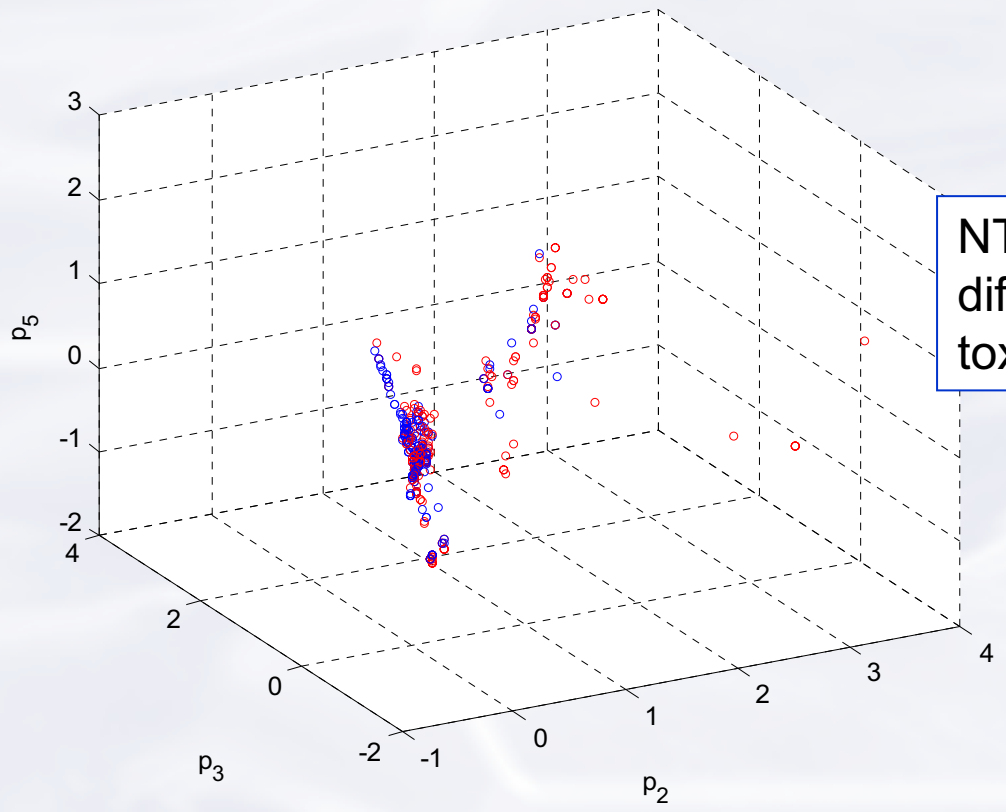
Principal Components – discrimination of NTP chemical space by carcinogenic potential



NTP space can be differentiated by toxicity endpoints.

Principal component scores: features (F=100 features selected by chi-sq)
red=Carc positive (p>0.5), blue=Carc negative (p<=0.5)

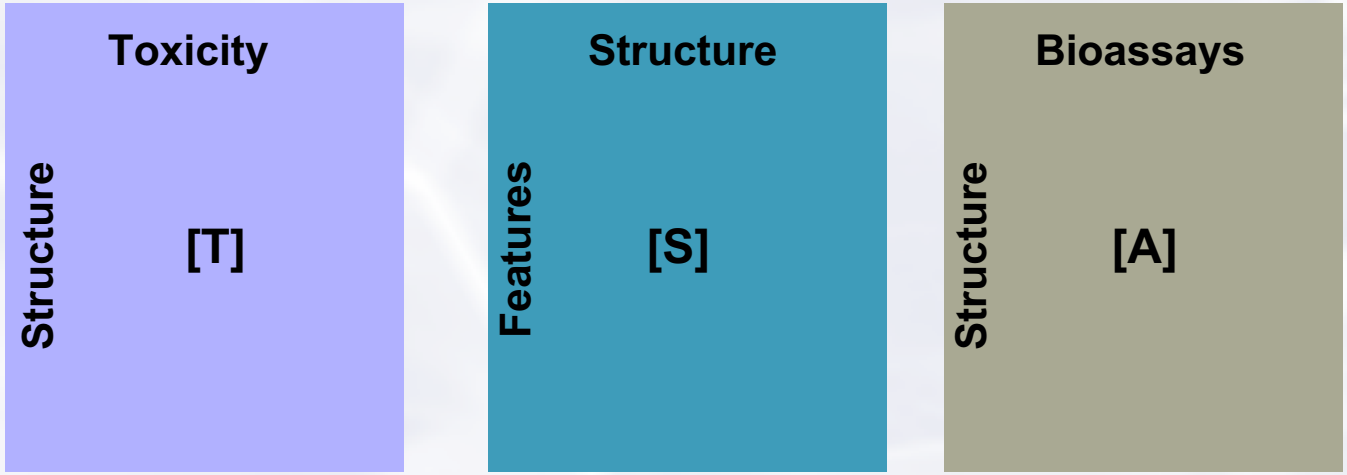
Principal Components – discrimination of NTP chemical space by acute toxicity



NTP space can be differentiated by toxicity endpoints.

Principal component scores: features (F=100 features selected by chi-sq)
(red=pLD50Rodent > -0.84, blue=pLD50Rodent <= -0.84)

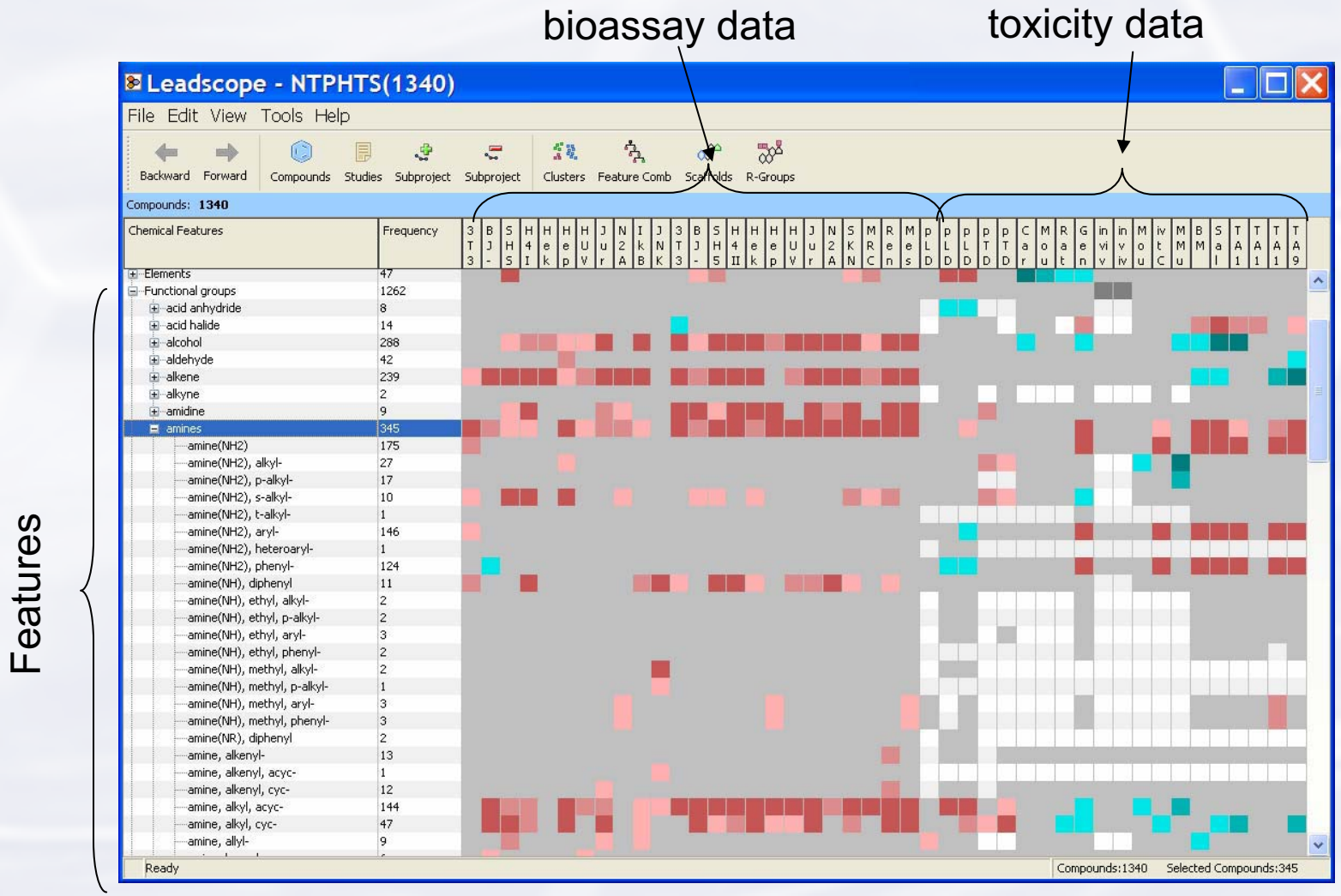
Linking bioassays to toxicity through structural features



Structure-Features matrix is the mathematical link between the [T] and [A] matrices.

This is my mathematical premise for this approach.

Very sparse chemical space for toxicity data



Structural features differentiating the profile

- 158 features are selected
- Z-scores are used to distinguish the activities

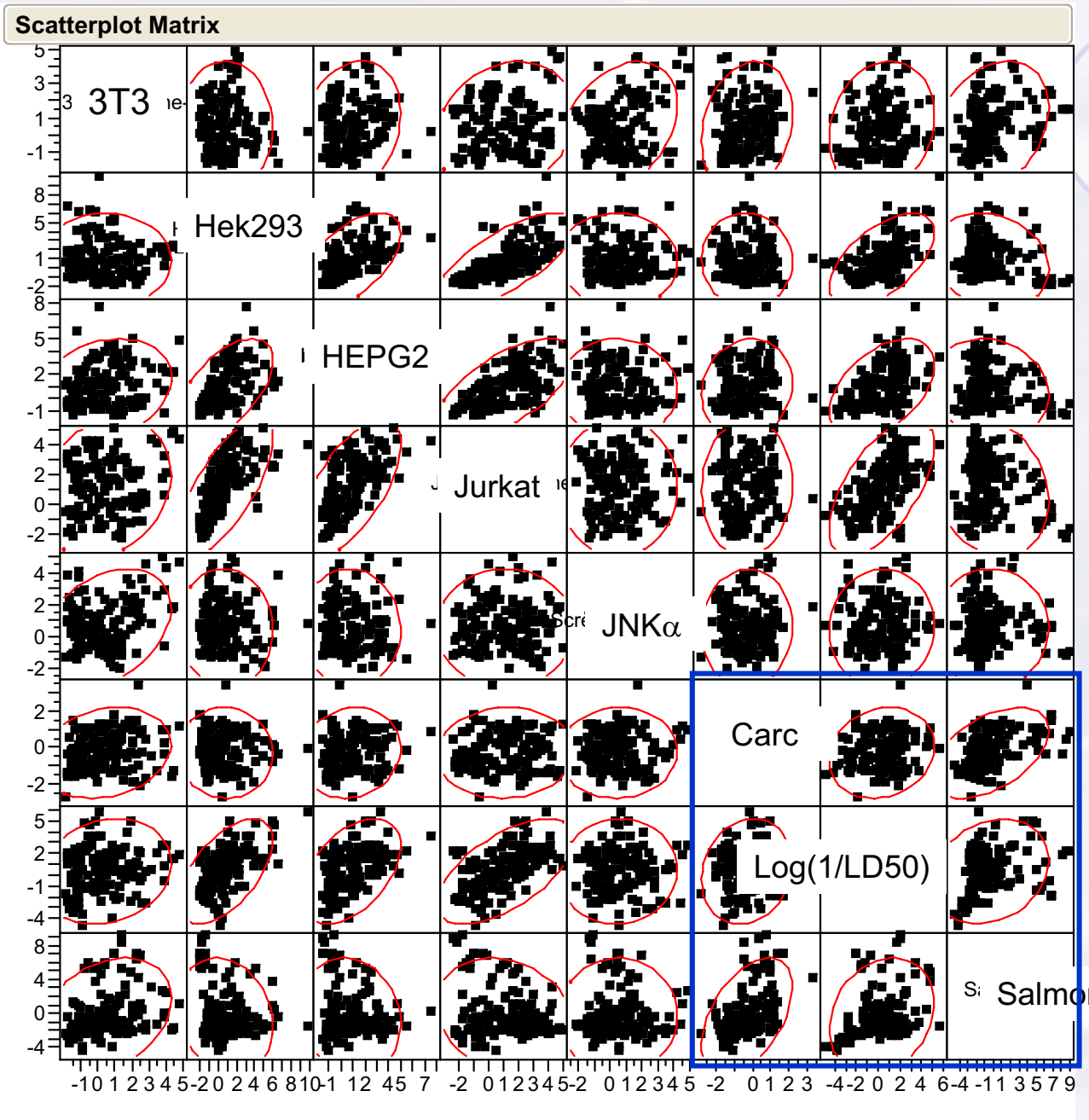
$$\text{z-score} = \frac{\text{class mean} - \text{overall mean}}{\text{standard error}}$$

- Multivariate analysis
  - Pearson correlations

$$r_{A_i T_j} = \frac{s_{A_i T_j}}{s_{A_i} s_{T_j}}$$

Examples of structural features

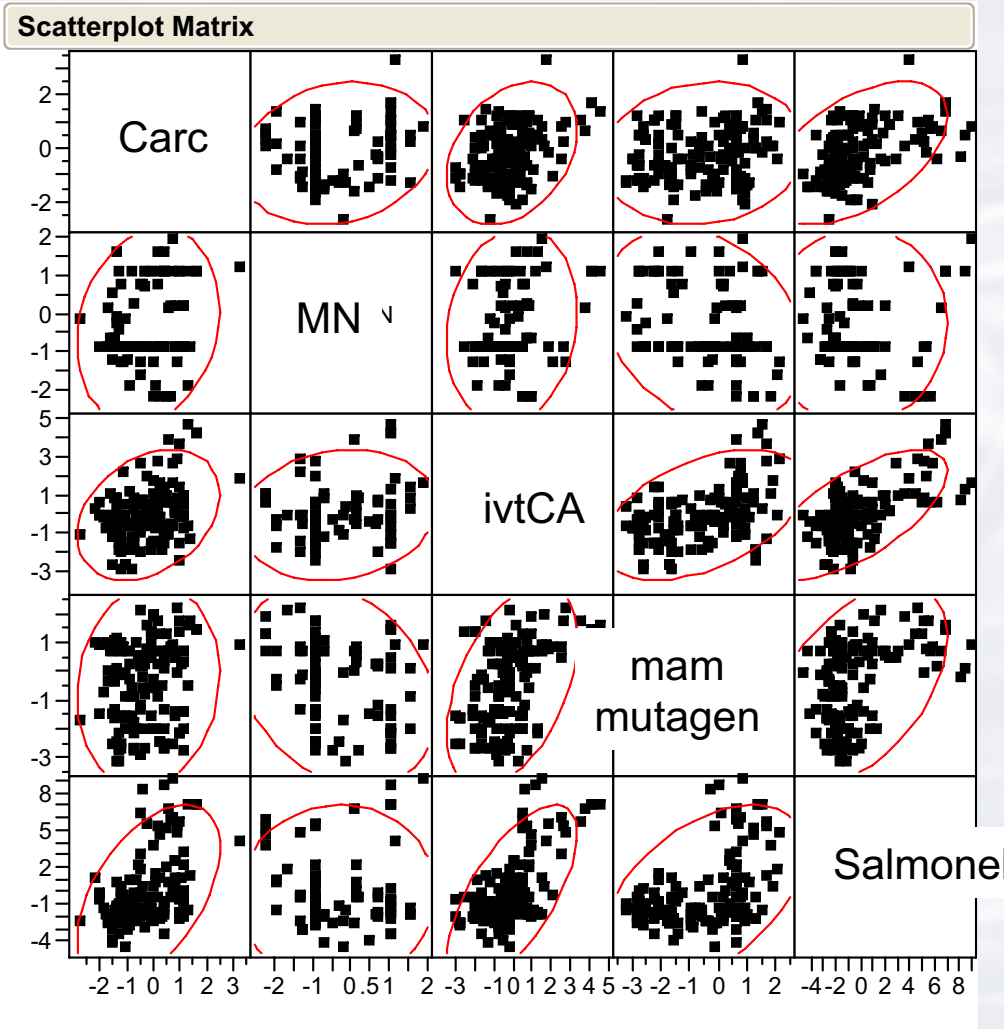
Chemical Features	%	3T3	Hek 293	Hep G2	JNK α	Jurkat	Carc _Call	pLD50
benzene, 1-R-,2-carbonyl-	7.7	2.81	-0.64	-0.10	4.46	-0.01	0.40	-2.78
benzene, 1-R-,2-alkoxy-	3.8	0.39	0.24	1.22	2.23	0.72	1.17	0.81
alcohol, alkyl, acyc-	8.4	-1.48	0.94	0.32	-1.23	0.84	-1.40	-2.80
alcohol, aryl-	11.9	4.36	2.17	2.04	1.56	4.58	-1.63	1.35
alcohol, cyclohexyl-	1.0	-0.39	0.32	-0.67	-0.05	0.99	-1.25	2.72
carbonyl, aminomethyl-	1.0	0.60	1.65	4.20	-1.47	2.81	0.43	1.63
halide, alkyl, acyc-	9.2	-0.53	1.93	1.22	0.48	1.26	0.66	3.35
steroid, 13-methyl-	1.9	0.94	5.15	2.61	-2.05	3.54	0.22	1.35



Same endpoints correlation as slide #8, but using the molecular descriptors. Now we can start seeing some patterns.

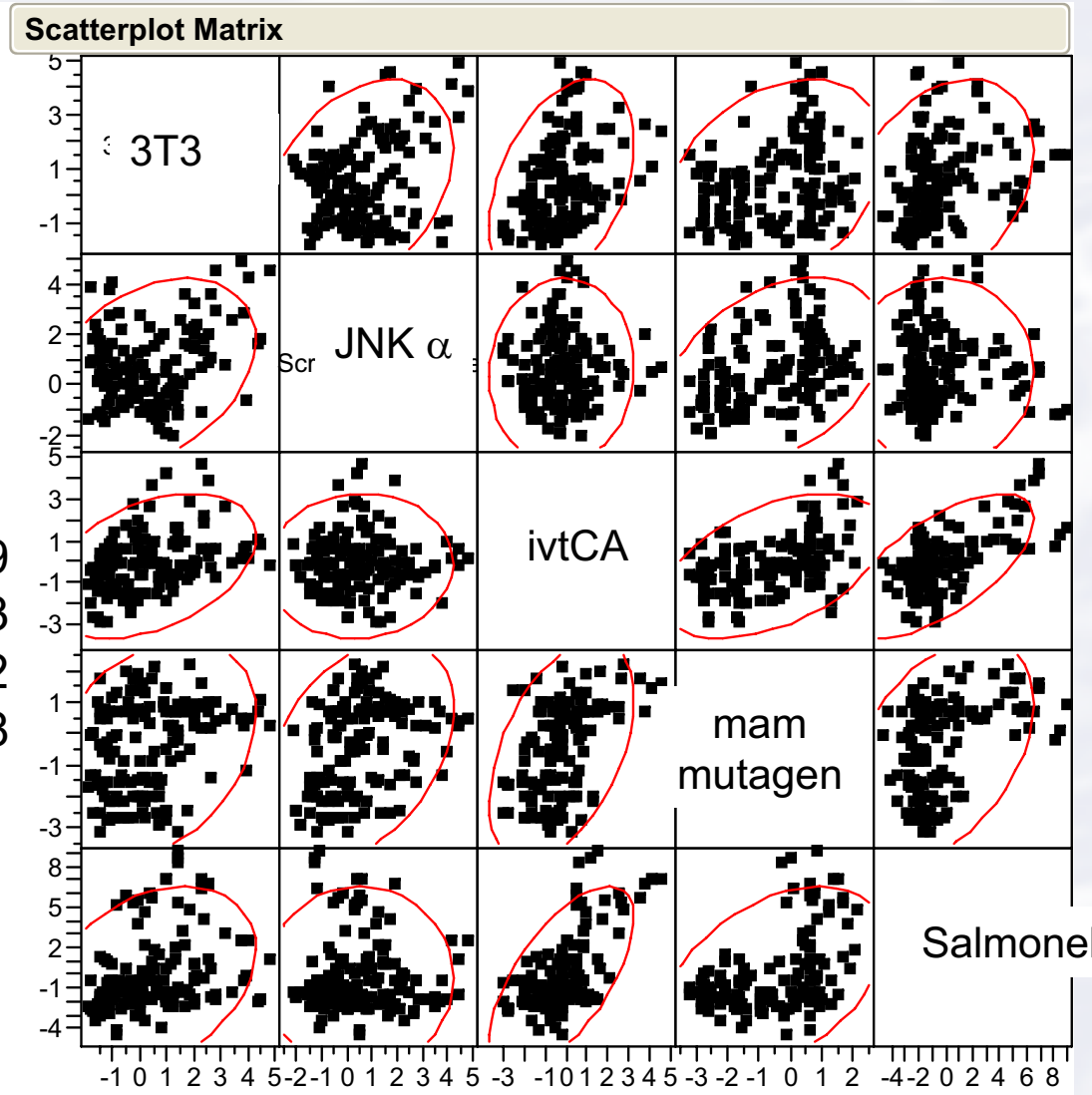
Correlation between genetic toxicity and carcinogenicity

Salmonella – Carc: 0.52
ivtCA – Carc: 0.32
Salmonella – MMut:0.49
ivtCA-MMut: 0.46
ivtCA-Salmonella: 0.65

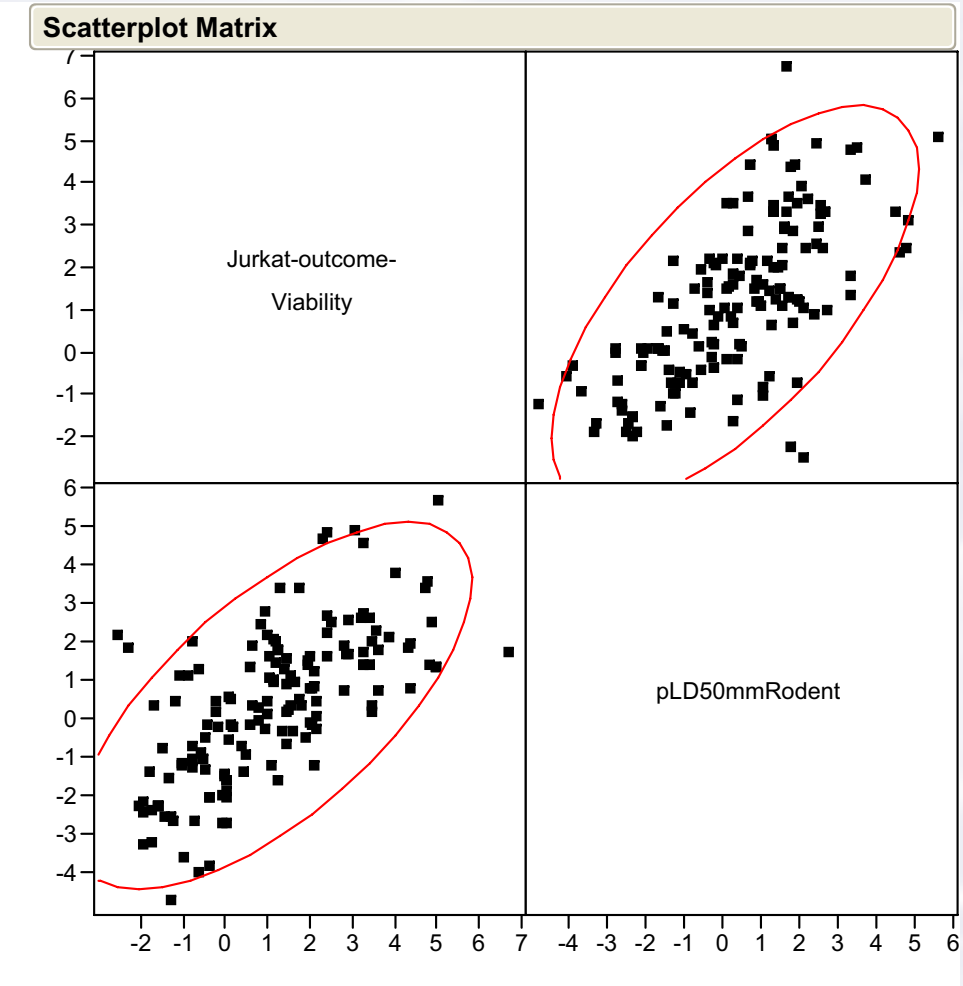


3T3 and JNK with mutagenesis…

ivtCA-3T3: 0.39
MMUT-3T3: 0.33
Salmonella-3T3: 0.32
MMut-JNKα: 0.38



Jurkat viability against Rodent LD50



Jurket-RodentpLD50: 0.68

Summary

- Correlations of bioassays and toxicity cannot be assessed at the compound level with the current toxicity database.
- Expanding structure dimensions to structural features or molecular descriptors, bioassays and toxicities can be mathematically correlated.
- Chemical selection (nomination) process for testing will benefit by incorporating chemoinformatics approaches.
- Further work is planned for gaining molecular level knowledge from these experiments.

Acknowledgement

- The Ohio State University
  - Professor J.F. Rathman
- Leadscope team