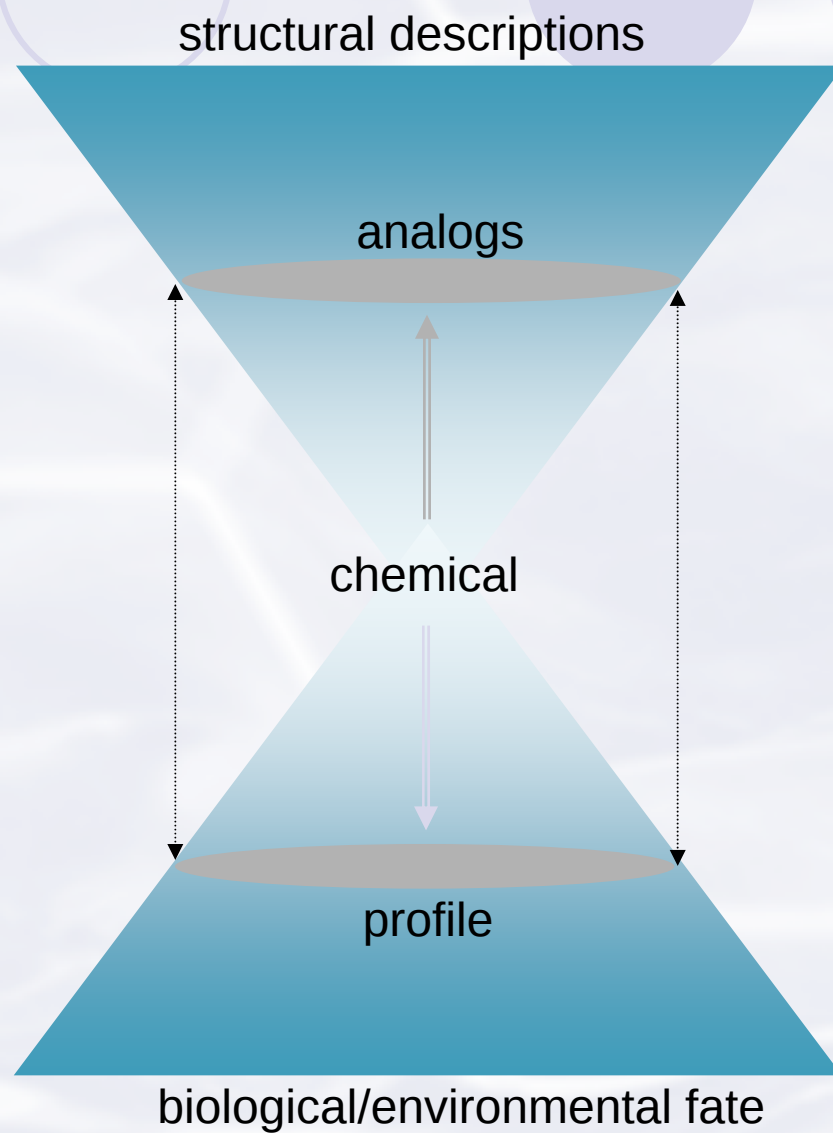


Applying data mining approaches to further understanding chemical effects on biological systems

Chihae Yang, Ann Richard, Jennifer Fostel

March 28, 2007

Linking biology to chemistry



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, H4Ile ,Hek293, HepG2, HUVEC, Jurkat, N2A
 - IκB signaling, JNK Alpha
 - Viability: Fred (7 strains), NCGC (13 strains)
 - 3T3, BJ, SHSY5, H4Ile ,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

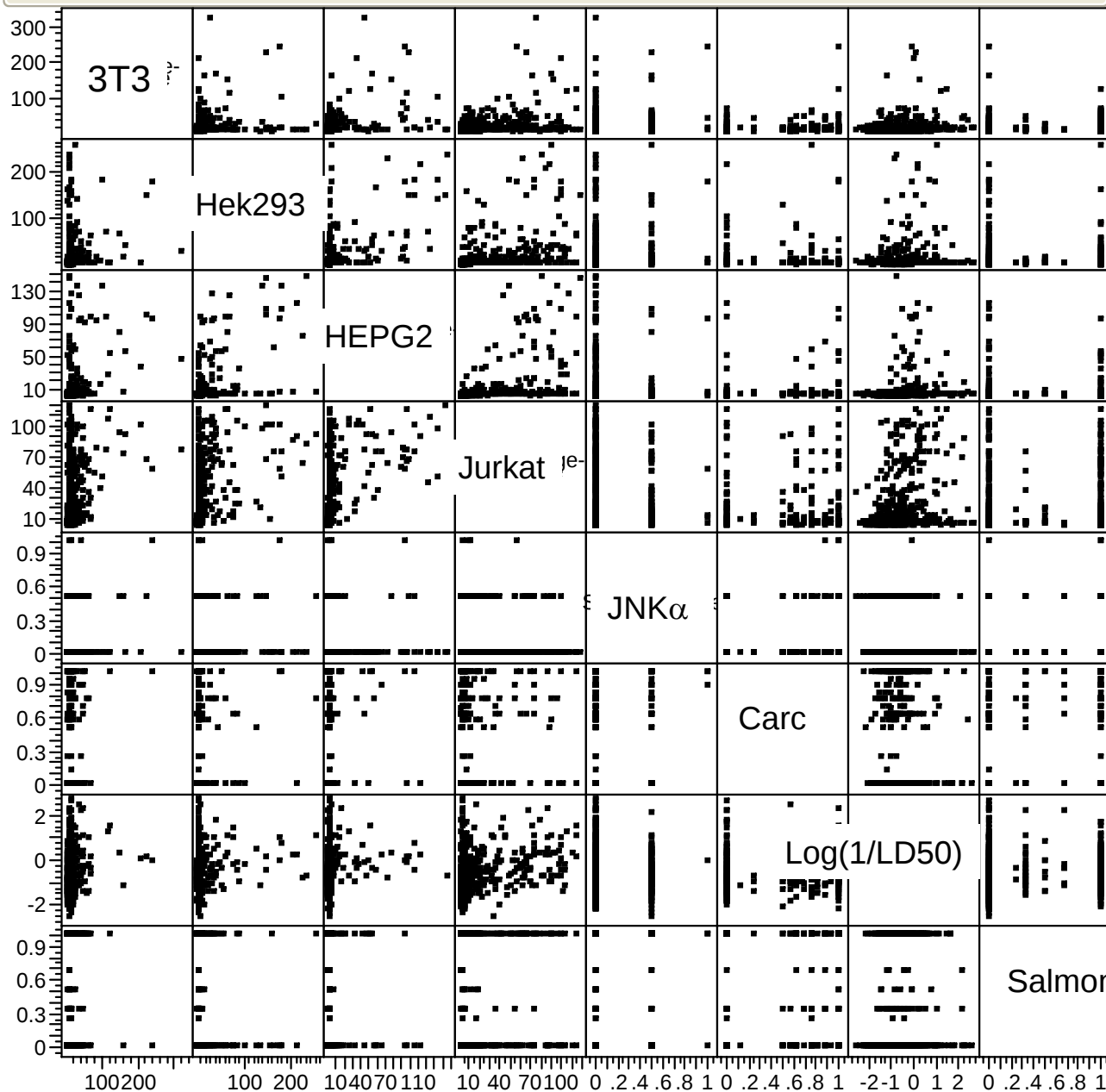
Sparse toxicity data space

		bioassay data																				toxicity data											
		Si	L	C	D	SH	3T3	H4II	BJ	Hek	Hep	HU	Jurk	N2f	SK	Jurk	MR	Hek	BJ	Hep	Mes	N2f	HU	FDA	NTP	FDA	CPD	CCR	NTP	Sam	RTE		
Compounds	N L15 #	#	5	14	1	3	4	1	11	5	7	8	8	14	10	12	9	12	11	19												RTE	
	N L15 #	#	6	12	2	3	6	2	8	4	7	10	10	14	8	8	12	20	13	16	NTP							CCR	S_GEN		RTE		
	N D15 #	#	8	14	1	4	4	3	20	4	7	16	20	12	12	24	19	25	19	26													
	N L15 #	#	5	12	2	3	2	2	12	8	4	14	10	10	9	16	10	9	9	27	NTP						CPD	CCR	NTP	Sam	RTE		
	NTP15 #	#	7	11	3	3	12	1	12	##	8	13	17	14	12	18	11	15	17	15													
	N L15 #	#	6	15	2	5	3	1	9	6	8	22	57	9	18	22	17	12	35	13	NTP						CPD	CCR	NTP	Sam	RTE		
	N L15 #	#	4	11	2	4	3	1	7	3	6	28	12	10	10	8	11	18	24	20											RTE		
	N L15 #	#	5	11	2	2	3	1	5	6	5	12	10	16	10	15	9	9	12	17											RTE		
	NTP16 #	#	5	14	13	6	7	2	12	9	9	38	6	17	86	33	58	##	##	37													
	N L16 #	#	7	13	2	5	3	1	9	3	11	16	8	12	14	24	18	23	15	27								CCR	S_GEN		RTE		
	NTP16 #	#	4	11	2	2	5	2	8	5	8	8	11	13	6	14	10	19	##	29													
	N L16 #	#	6	10	2	3	3	1	5	3	6	14	9	12	11	8	16	14	16	26								CCR	S_GEN		RTE		
	NTP16 #	#	3	7	2	3	4	2	6	4	4	18	7	9	9	20	7	11	8	13													
	N L16 #	#	7	11	3	4	5	2	9	4	9	13	11	11	14	20	13	28	34	35								CPD	CCR	S_G	Sam	RTE	
	N L16 #	#	6	12	2	5	3	2	7	5	7	15	7	10	10	13	19	17	13	29	NTP							CPD	CCR	NTP	Sam	RTE	
	N L16 #	#	5	12	2	2	4	2	7	6	8	11	8	8	12	15	6	10	14	19												RTE	
	NTP16 #	#	7	10	3	3	4	1	9	4	8	12	8	8	11	9	8	7	16	19													
	N L16 #	#	6	10	2	3	4	1	6	4	3	10	10	8	11	14	7	6	12	19	NTP								CCR	S_GEN		RTE	
	N L16 #	#	26	14	5	5	4	1	8	7	13	20	31	10	11	31	10	19	##	21	NTP												
	N L16 #	#	5	25	2	3	2	1	11	3	5	15	12	13	27	15	10	10	11	24	FDA						FDA		CARC				
	N L17 #	#	3	10	2	6	3	1	10	5	6	20	6	10	7	19	12	10	17	23	NTP								CPD	CCR	NTP	SampleCa	
	N L17 #	#	4	19	30	3	4	2	8	3	9	16	14	11	10	17	22	11	30	23	NTP												
	N L17 #	#	4	17	3	3	5	2	10	3	8	6	15	11	10	15	15	6	20	25	NTP								CCR	S_GEN		RTE	
	NTP17 #	#	7	10	4	5	5	3	11	5	6	18	13	11	15	35	20	20	15	36													
	N L17 #	#	3	12	2	3	2	1	8	4	9	25	7	11	6	10	9	9	17	18	NTP								CCR	S_GEN		RTE	
	N L17 #	#	7	13	1	2	7	1	10	5	6	10	9	12	10	15	10	15	13	17													RTE
	N L17 #	#	4	10	2	3	3	2	8	5	4	13	9	10	8	10	9	14	12	29													RTE
	N L17 #	#	5	13	2	3	6	2	8	7	4	13	12	15	6	21	13	14	9	21													
	N L17 #	#	5	14	2	2	5	2	8	3	4	11	10	6	11	16	13	14	11	36										CCR	S_GEN		
	N L17 #	#	6	14	2	4	6	2	6	5	7	10	9	8	10	11	12	13	14	28									CCR	S_GEN			
	N L17 #	#	4	10	2	3	3	2	7	4	6	18	6	16	9	8	14	19	15	26	NTP								CPD	CCR	NTP	Sam	RTE
	N L18 #	#	3	10	2	4	3	2	9	2	6	9	12	7	12	14	6	18	30	19													
	N L18 #	#	7	12	2	4	4	1	6	4	5	16	12	11	11	12	12	12	21	20	NTP								CCR	S_GEN		RTE	
	N L18 #	#	8	11	2	5	3	2	9	5	9	30	20	15	11	34	30	21	27	39	NTP												RTE
	NTP18 #	#	4	9	2	3	5	2	13	5	5	13	6	10	14	12	13	11	22	20													
	NTP18 #	#	5	10	2	4	4	1	8	3	6	12	11	7	15	8	13	9	14	14													
	N L19 #	#	4	6	2	5	5	4	6	4	4	7	6	11	11	9	8	18	10	11	NTP								CPD	CCR	S_G	Sam	RTE
	N L19 #	#	6	11	2	2	4	2	6	6	6	19	8	10	12	15	10	13	12	19	NTP								CPDB	IS	NTP	Sam	RTE

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.

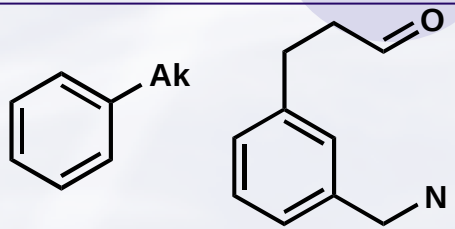
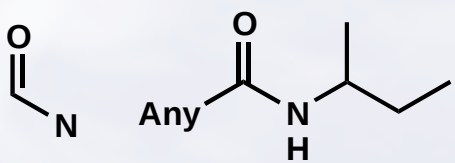
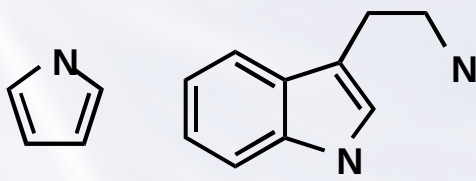
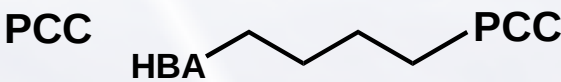
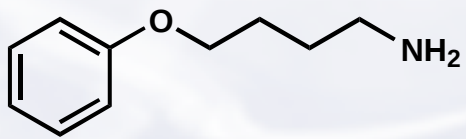
Scatterplot Matrix



multivariate correlation

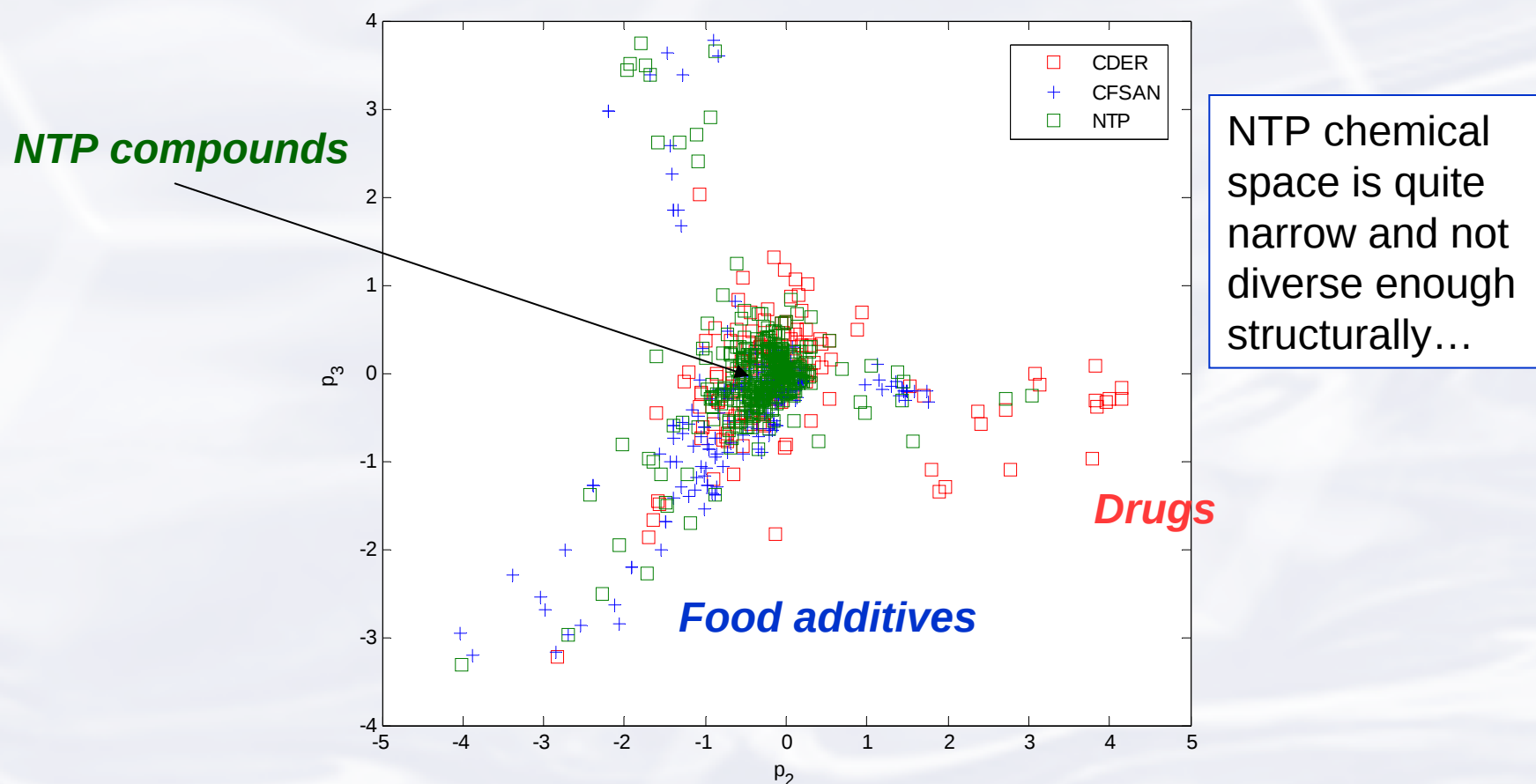
At compound level,
virtually not possible
to correlate any
activities or toxicity
data.

Representing structures with Leadscope molecular descriptors

Benzenes	
Functional groups	
Heterocycles	
Pharmacophores	
Spacers	
User defined features	knowledge addition

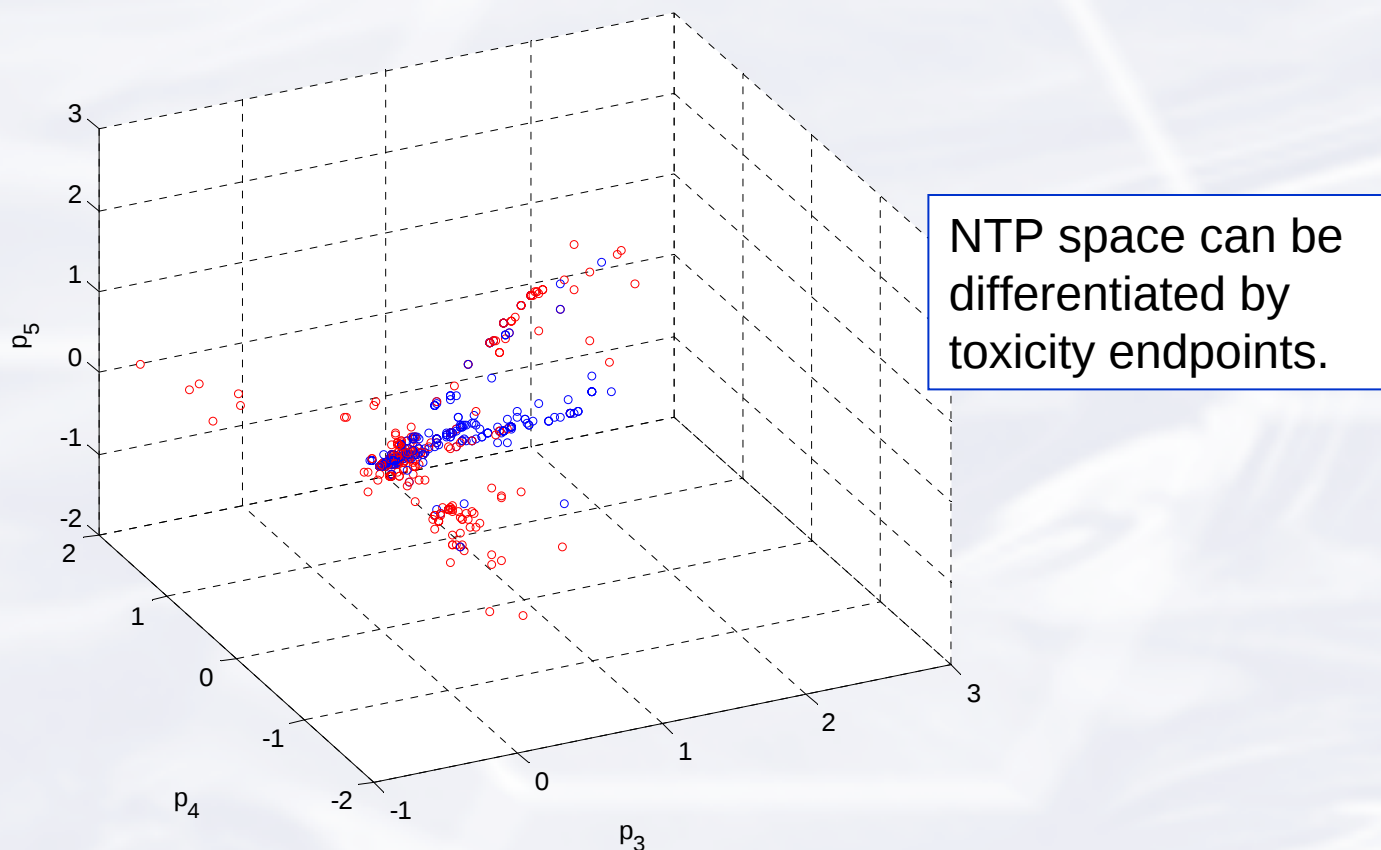
Applying a set of molecular descriptors to expand chemistry space.

Structure domain characterization by principal component analysis



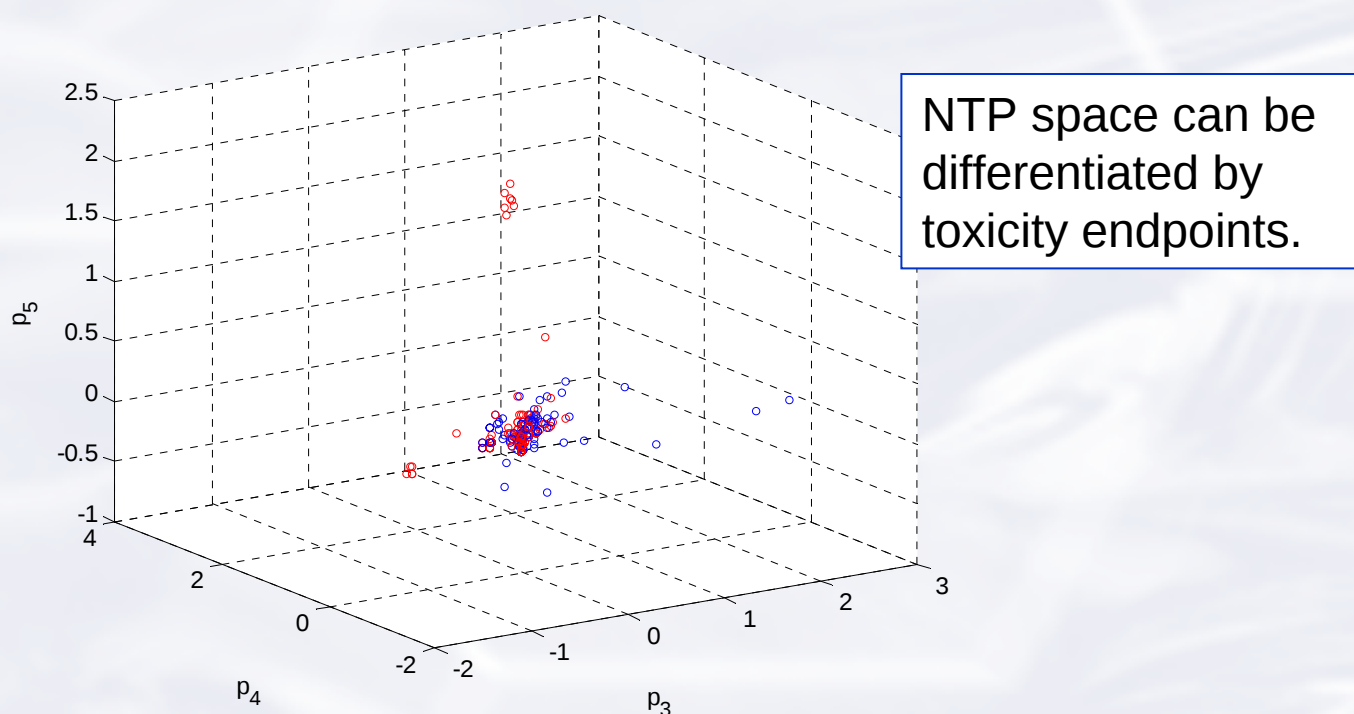
using selected features present in >10 compounds

Principal Components – discrimination of NTP chemical space by mutagenic potential



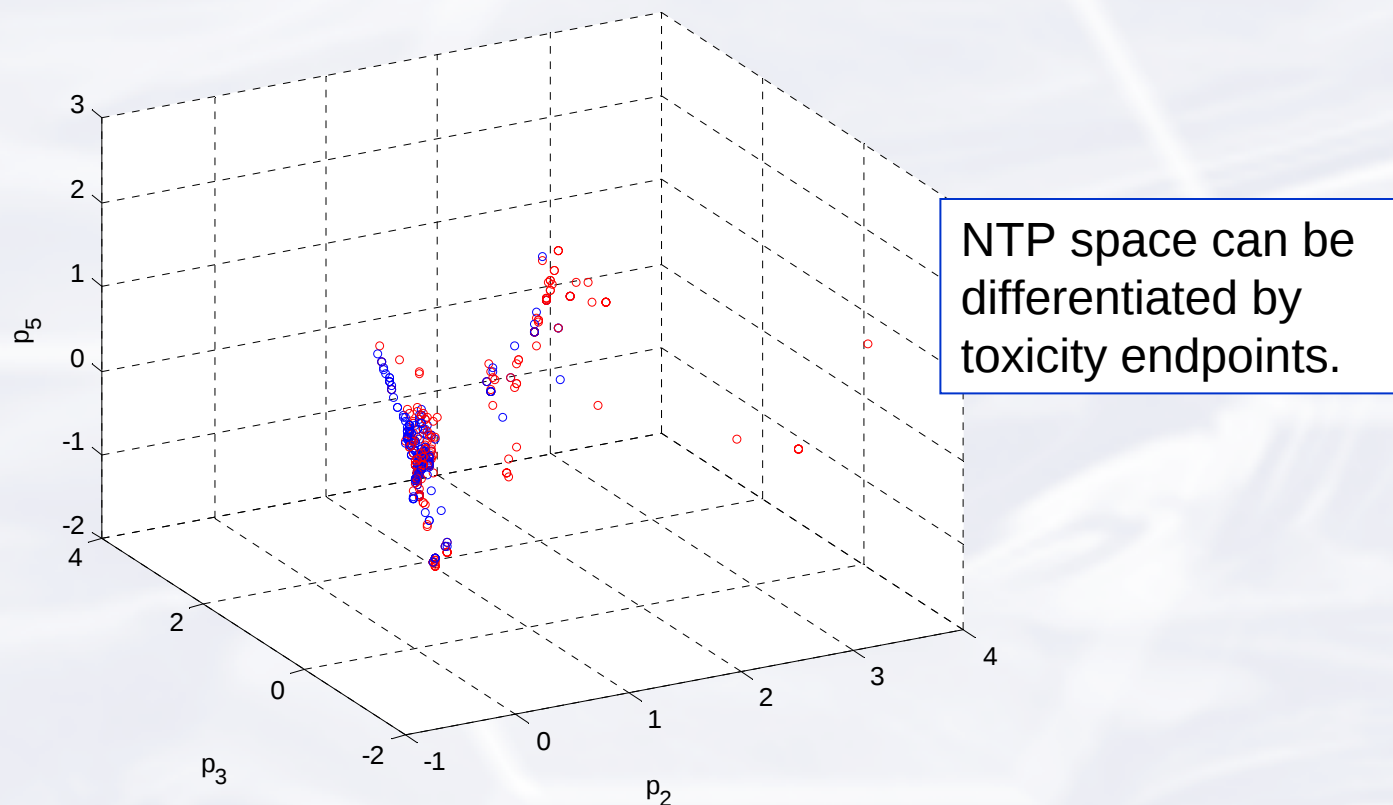
Principal component scores: features (F=100 features selected by chi-sq)
red=Salmonella positive ($p > 0.5$), blue=Salmonella negative ($p \leq 0.5$)

Principal Components – discrimination of NTP chemical space by carcinogenic potential



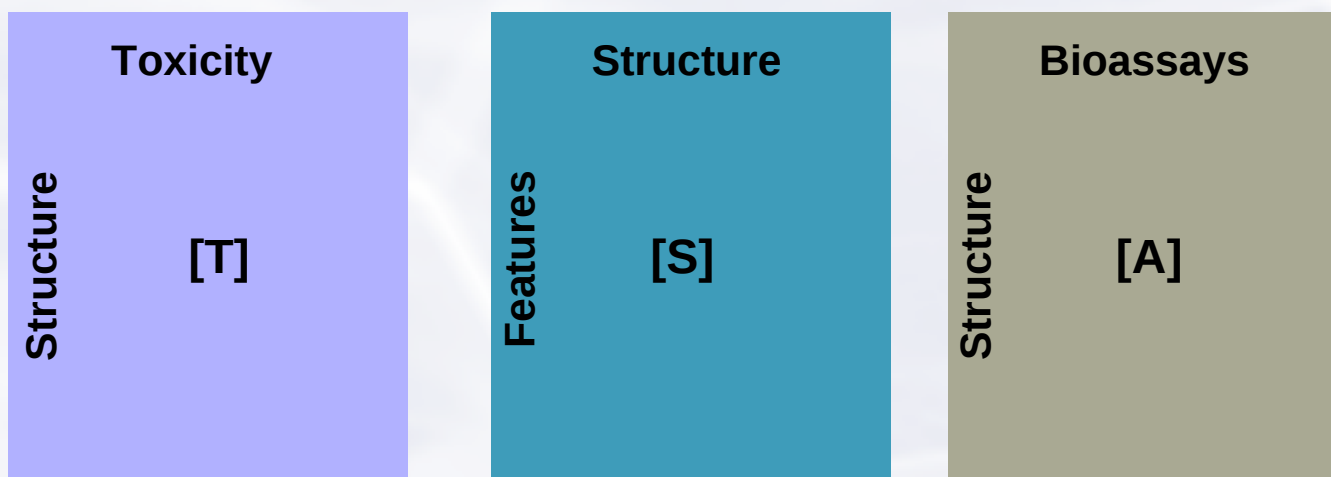
Principal component scores: features (F=100 features selected by chi-sq)
red=Carc positive ($p > 0.5$), blue=Carc negative ($p \leq 0.5$)

Principal Components – discrimination of NTP chemical space by acute toxicity



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

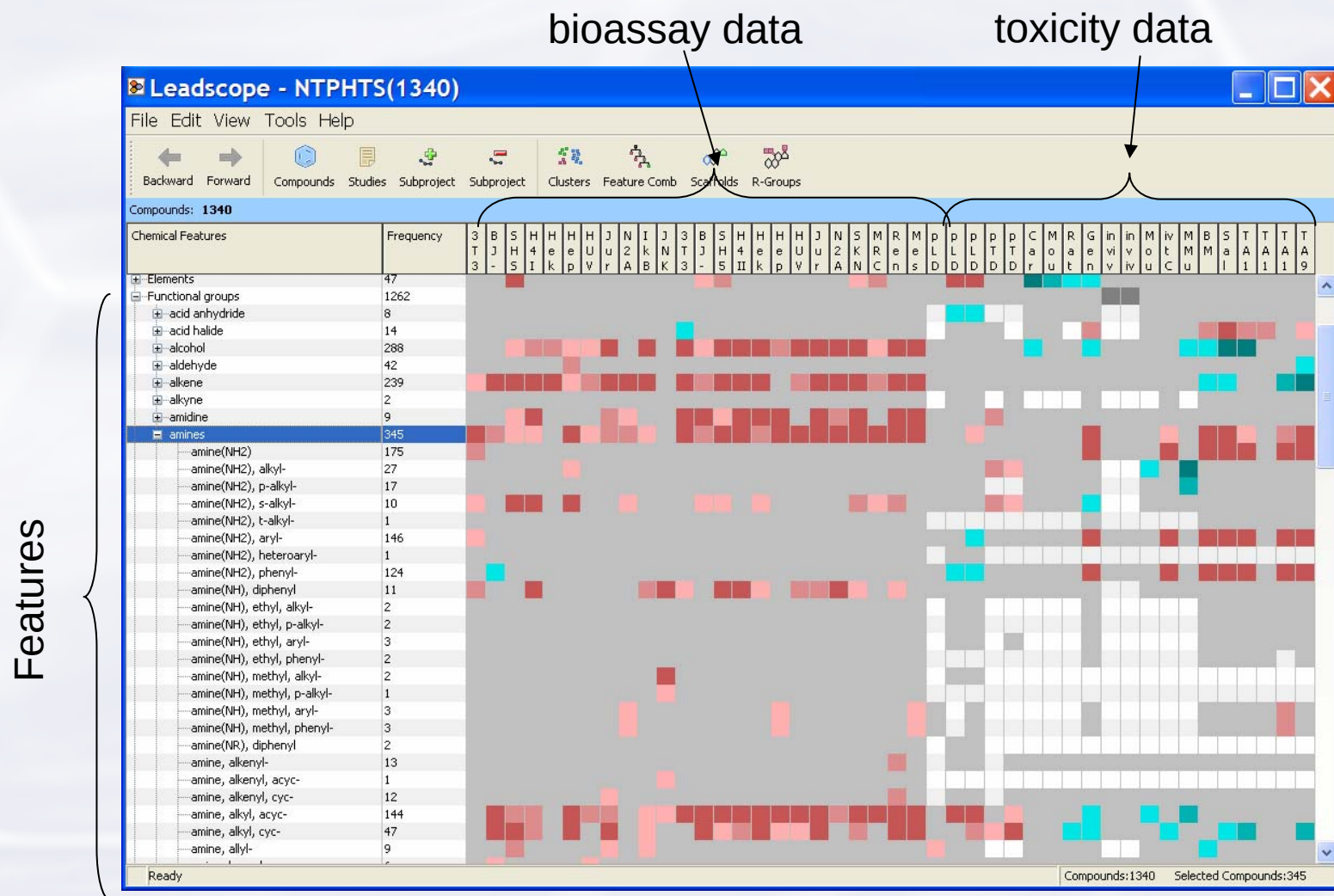
Linking bioassays to toxicity through structural features



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

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

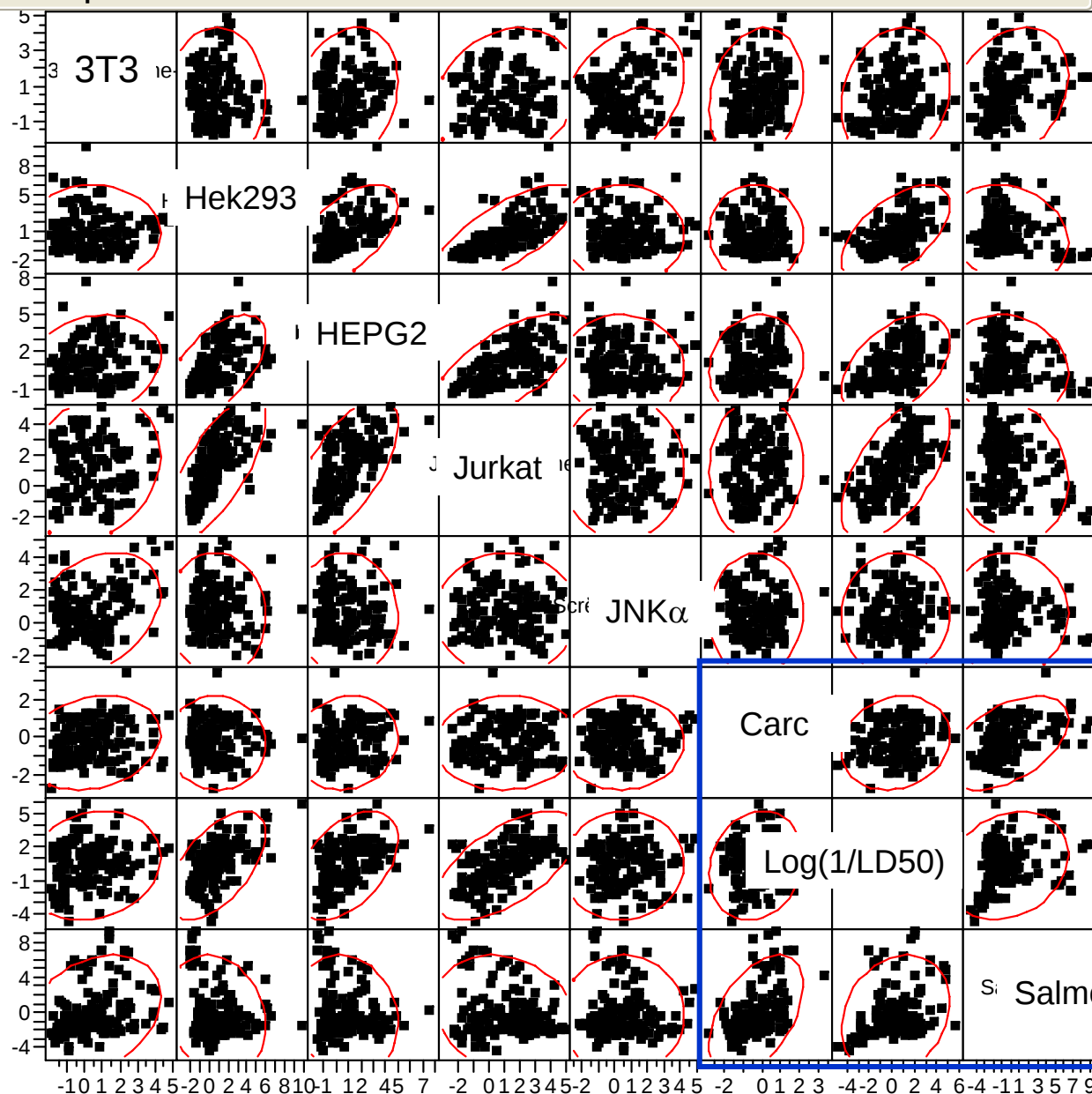
- Multivariate analysis
 - Pearson correlations

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

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

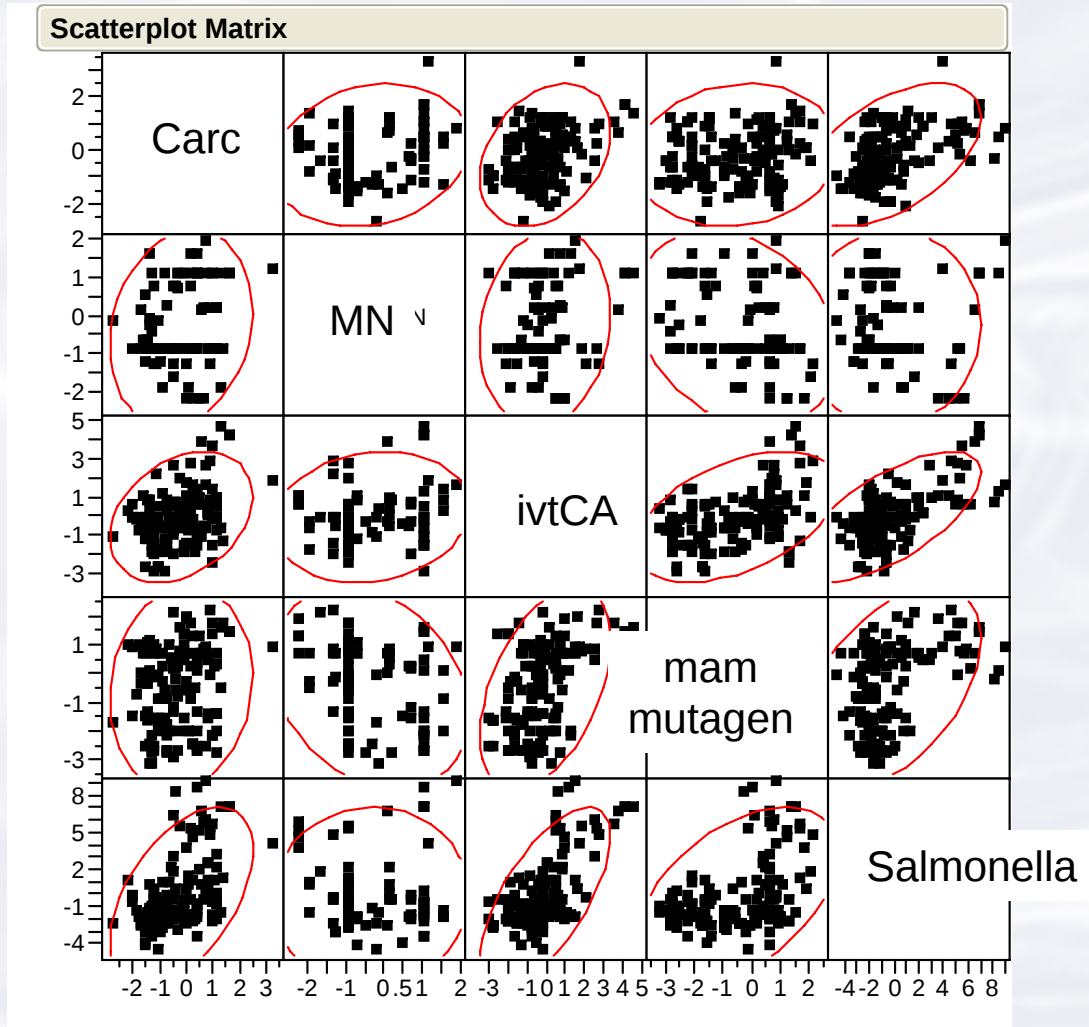
Scatterplot Matrix



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

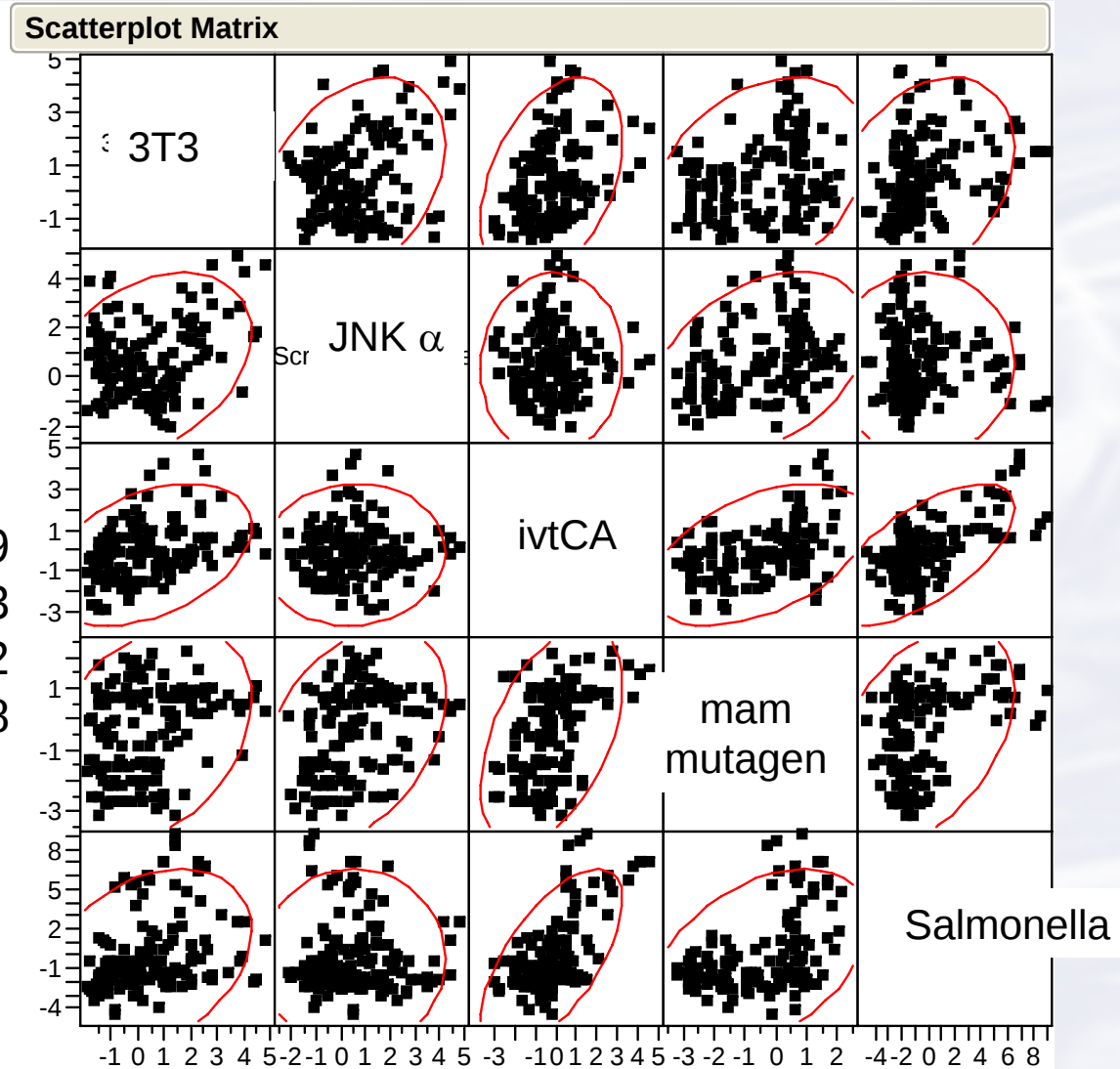
Correlation between genetic toxicity and carcinogenicity

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

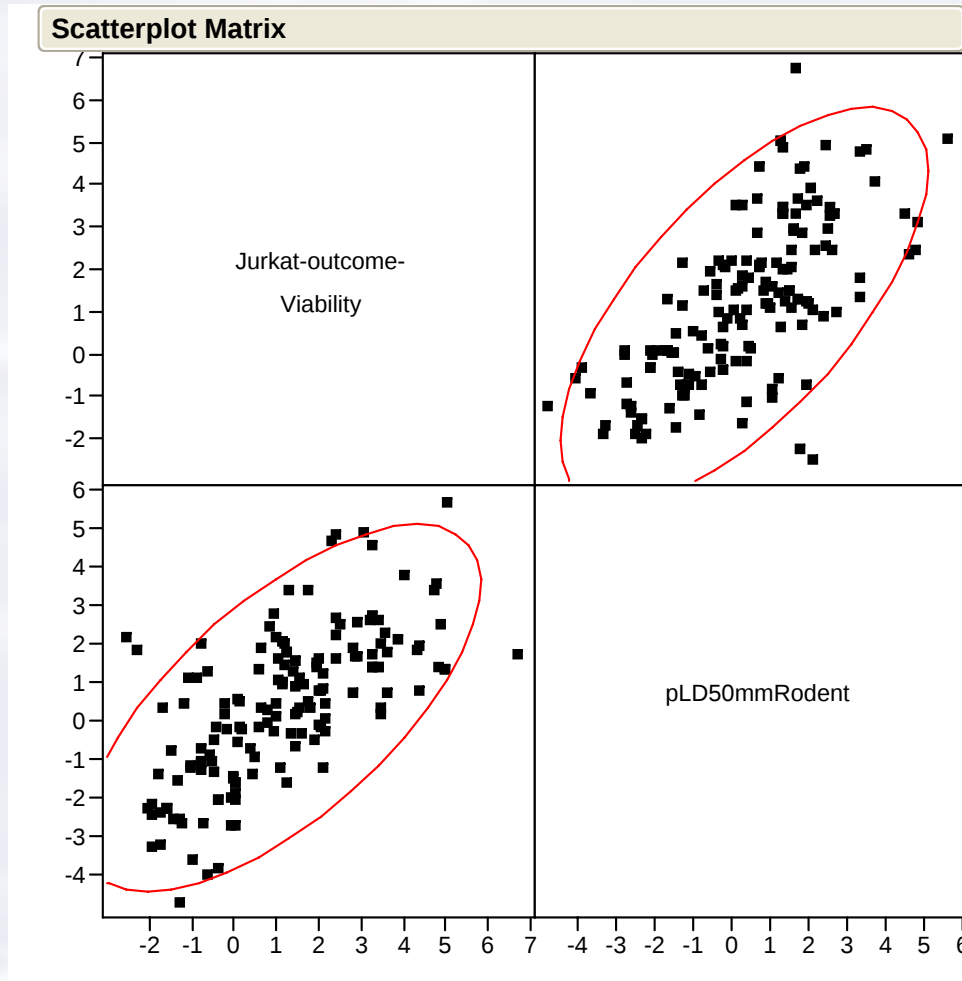


3T3 and JNK with mutagenesis...

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



Jurkat viability against Rodent LD50



Jurket-RodentpLD50: 0.68

Summary

A decorative graphic at the top of the slide features five circles arranged horizontally. The first, third, and fifth circles are solid light purple, while the second and fourth are hollow with a light purple outline. Behind these circles and across the top of the slide are faint, light blue wavy lines.

- 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