Genomic Signatures of Carcinogenicity

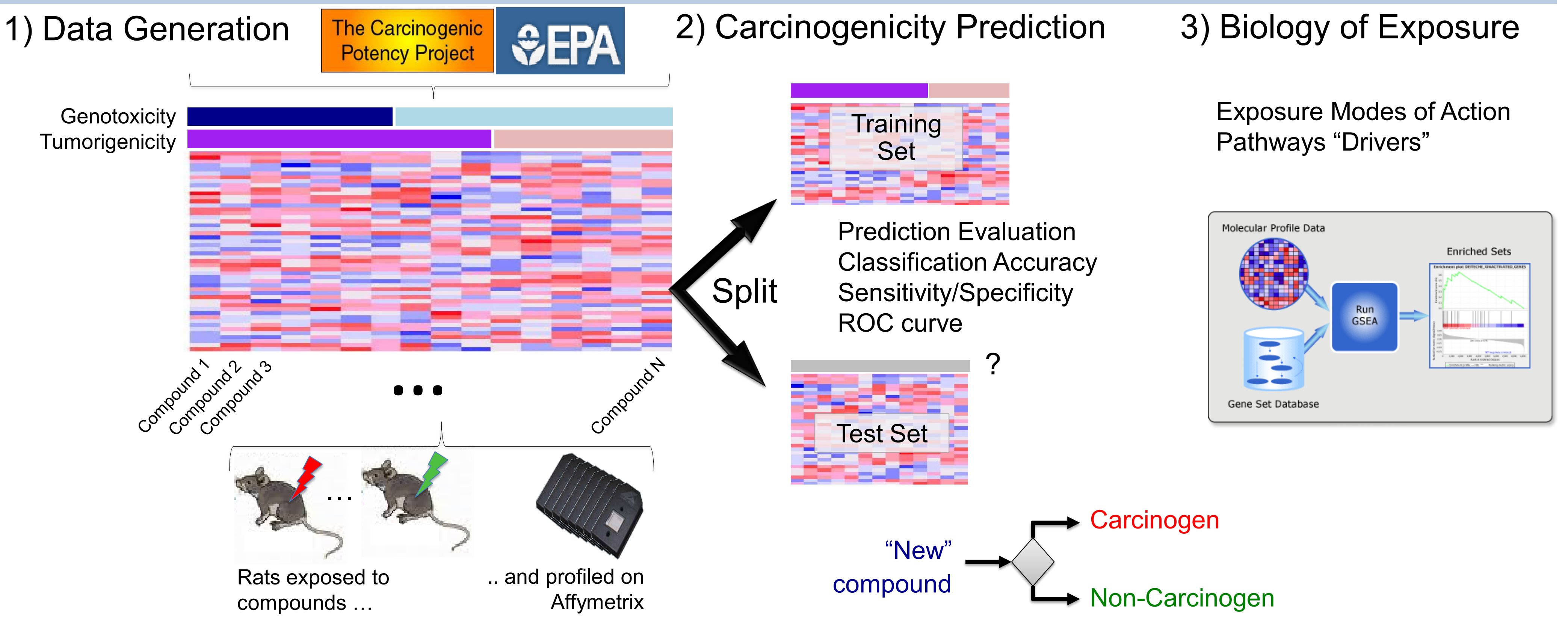
Daniel Gusenleitner, Tisha Melia, Harold Gómez, David Sherr, Stefano Monti
Department of Bioinformatics, Boston University, 24 Cummington Street, Boston, MA 02215

Overview

There are around 80,000 chemical compounds used in industry, many of which are suspected carcinogens. Standard approaches to carcinogen testing are costly and time-consuming and, as a result, only approximately 1,500 of the chemicals in commercial use have been tested. Additionally, some chemicals can have synergistic effects, making the characterization of carcinogenic compounds even more difficult as combinations have to be considered.

The goal of this project is the development of computational models of carcinogenicity based on gene expression, to classify the carcinogenic potential of individual or complex mixtures of environmental pollutants and/or therapeutics, and to study their mechanisms of action. To this end, we analyzed the **DrugMatrix** dataset, a large collection of 3610 gene expression microarray profiles from rats treated with 188 well-characterized chemicals, including genotoxic and non-genotoxic carcinogens, as well as non-carcinogens.

	Liver	Cell Culture	Kidney	Heart	Thigh Muscle	All Tissues
All samples	1380	580	902	641	107	3610
Untreated	279	113	335	231	36	994
Genotoxic	233	123	157	324	0	599
Non-Genotoxic	868	344	410	86	71	2017
Carcinogenic	691	276	308	195	35	1505
Non-Carcinogenic	410	191	259	215	36	1111
# Chemicals	110	69	71	56	12	188



1) Data Generation

The Carcinogenic Potency Project

EPA

Genotoxicity
Tumorigenicity

Compound 1
Compound 2
Compound 3
...
Compound N

Rats exposed to compounds …

.. and profiled on Affymetrix

2) Carcinogenicity Prediction

Split

Training Set

Prediction Evaluation
Classification Accuracy
Sensitivity/Specificity
ROC curve

Test Set

?

"New" compound

Carcinogen

Non-Carcinogen

3) Biology of Exposure

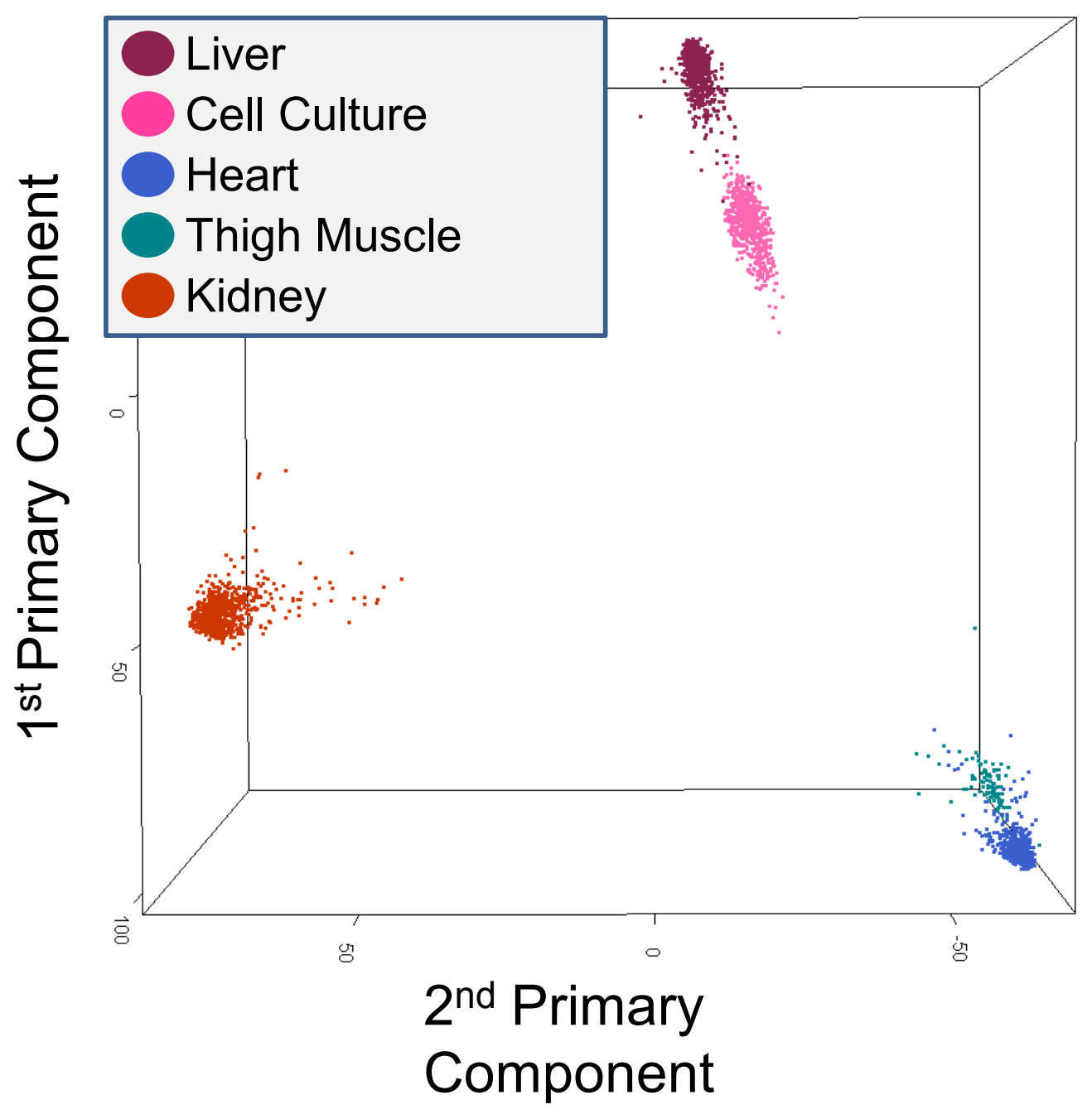
Exposure Modes of Action
Pathways "Drivers"

Genotoxicity and Carcinogenicity

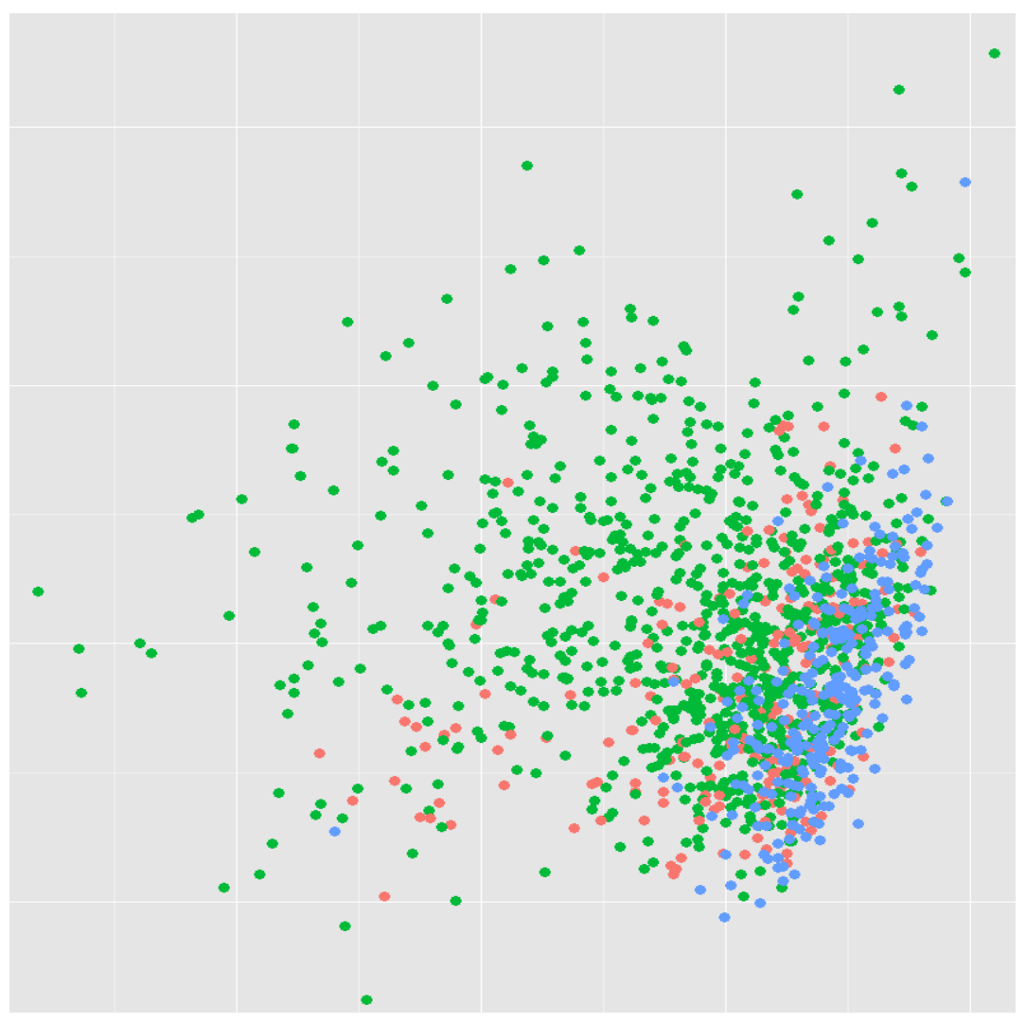
Genotoxicity is defined as the property of being damaging to DNA, thereby being capable of causing mutations and potentially cancer. Genotoxicity can be assessed by relatively simple tests, such as the Ames and Salmonella tests, which have moderately good sensitivity and specificity. *Carcinogenicity* is the property of being cancer-causing and while most carcinogens are also genotoxic, some are not. Current methods for testing carcinogenicity are based on the 2-year rodent bioassay, which is a very expensive and time-consuming testing protocol. Novel, cost-effective and accurate testing methods are therefore needed.

Exploratory Data Projection

Tissue Type

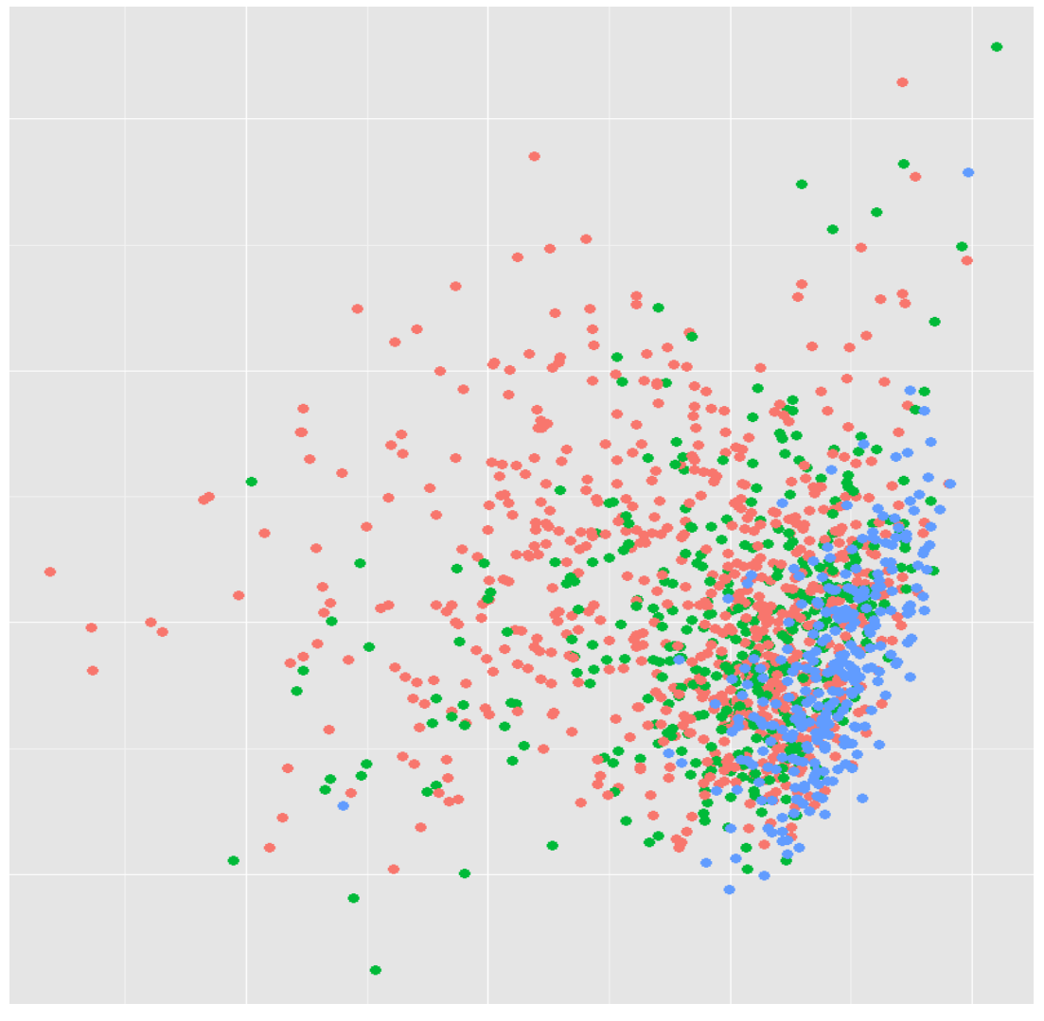


Genotoxicity



● Untreated
● Non-Genotoxic
● Genotoxic

Carcinogenicity

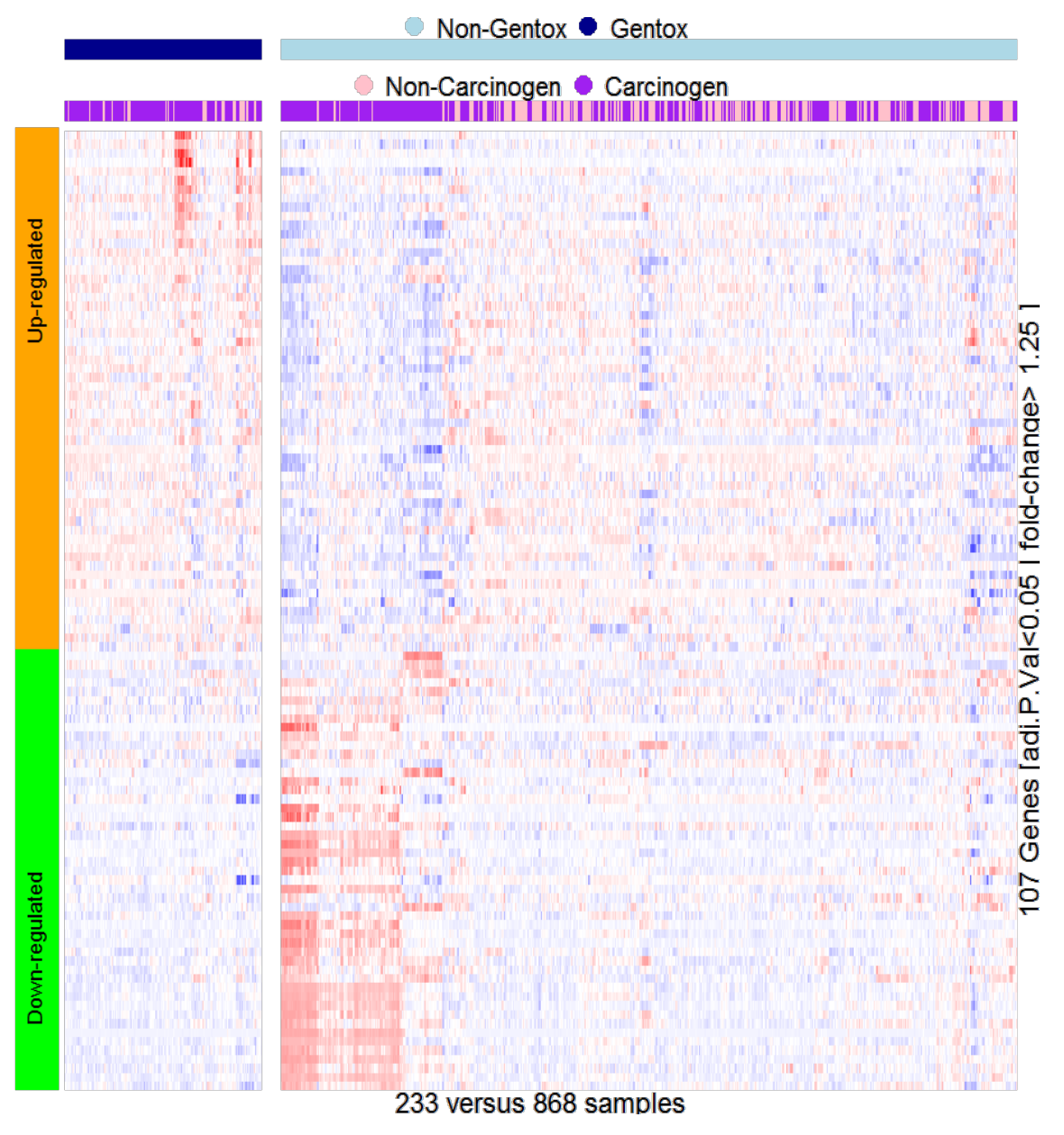


● Untreated
● Non-Carcinogenic
● Carcinogenic

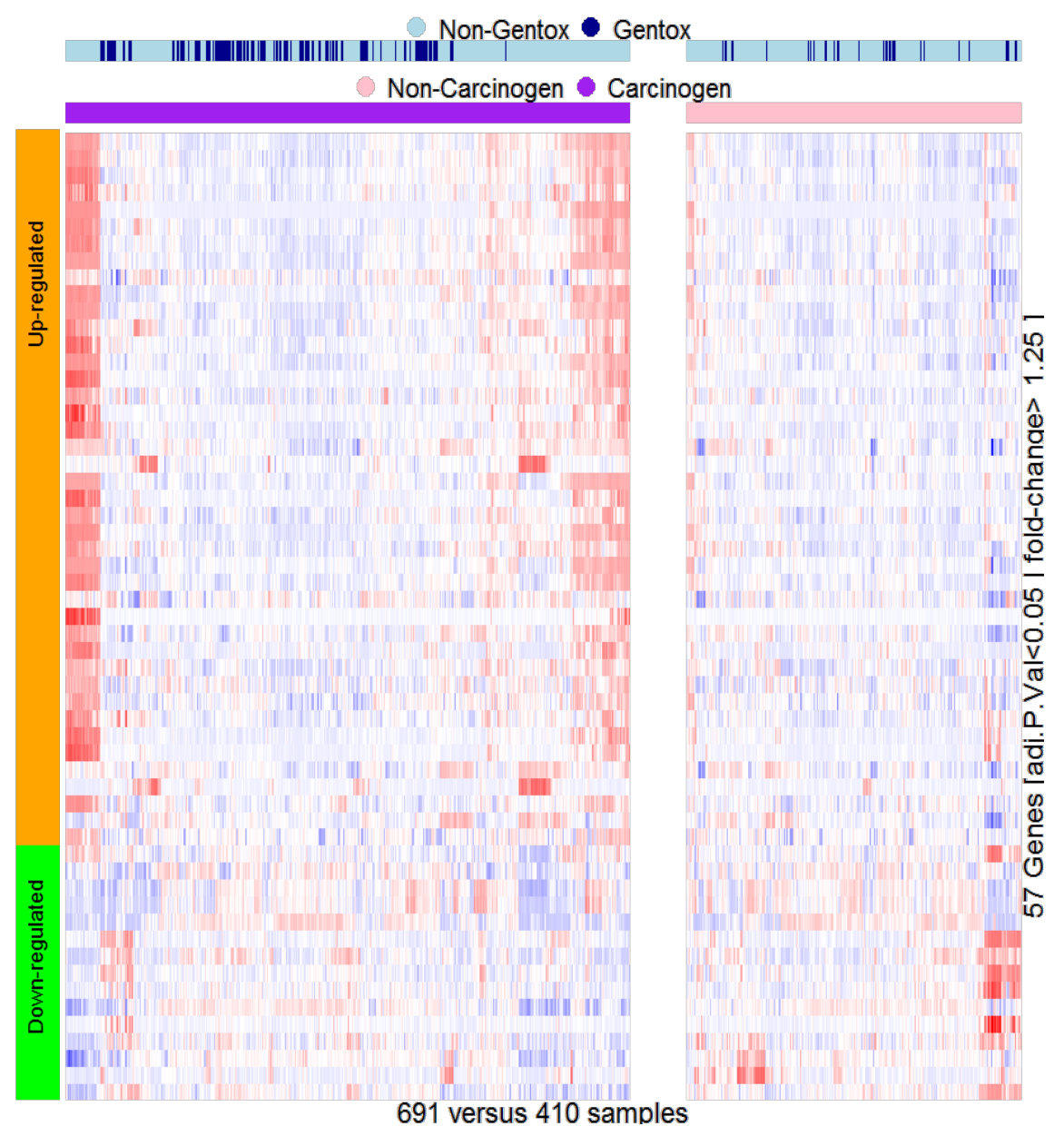
Comparative Marker Selection

Identifies 100s of transcripts differentially expressed ("signatures")

Liver:
Genotoxic vs. Non-Genotoxic



Liver:
Carcinogen vs. Non-Carcinogen



Number of Biomarkers

Tissue	Genotoxicity +	Genotoxicity –	Carcinogenicity +	Carcinogenicity –
Liver	58	49	42	15
Kidney	32	25	13	7
Cell Culture	172	313	61	33
Heart	21	27	5	9
All	109	80	348	197
	FDR≤0.05, Fold-Change≥1.25			

● Genotoxic ● Non-Genotoxic ● Carcinogenic ● Non-Carcinogenic

Signatures Annotation

Differential signatures were annotated by enrichment analysis, whereby genesets representing pathways and transcription factor targets were tested for "over-representation" in a phenotype's signatures.

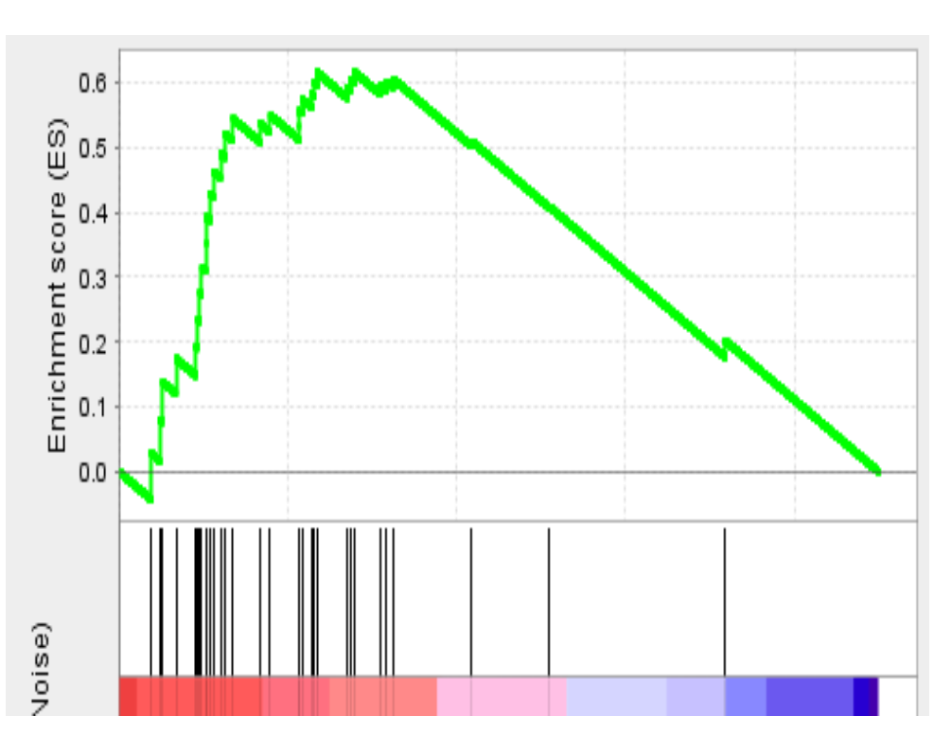
Pathways Enriched

Genotoxicity
- ➢ DNA-damage response, Cell Cycle progression, Metabolism, Apoptosis.

Carcinogenicity
- ➢ Fatty acid oxidation, Metabolism, Metabolism of Xenobiotics, Immune response, Regenerative proliferation.

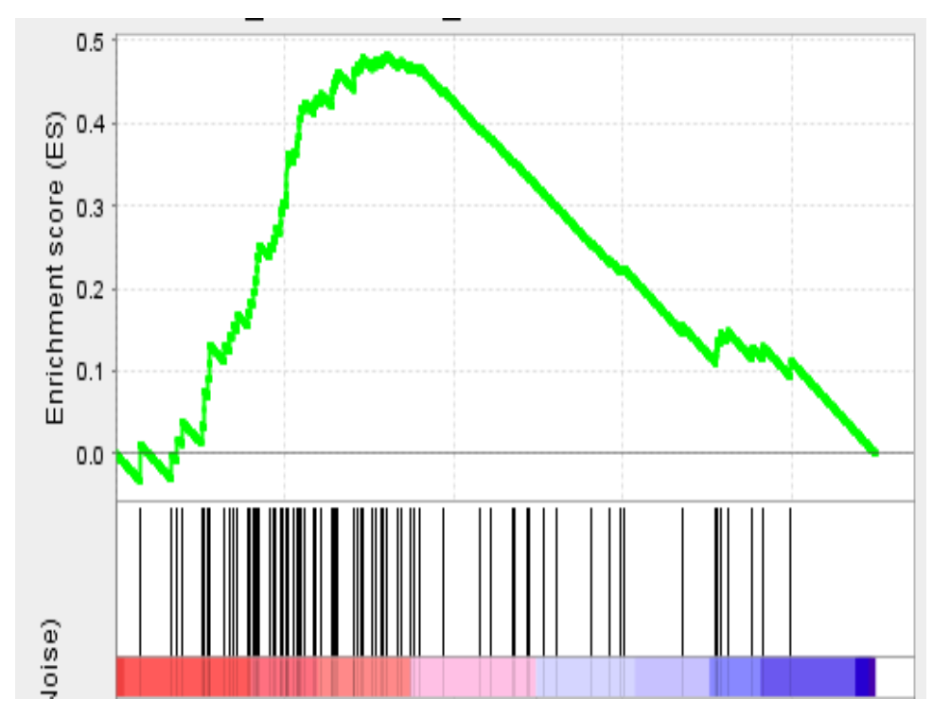
Gene Set Enrichment Analysis

Cell-culture:
p53 independent damage response



Genotoxic vs. Non-Genotoxic

Liver:
Oxidative phosphorylation



Carcinogen vs. Non-Carcinogen

Predicting Genotoxicity and Carcinogenicity

Gene expression-based prediction

We used transcript expression values as predictors. A candidate set of the 5000 most varying transcripts was selected to train and test Random Forests, an ensemble classifier well suited for large genomic datasets. Classifier training was repeatedly performed on 70% of the data and tested on the remaining 30% split to obtain unbiased prediction estimates. Classification thresholds were calibrated on the training set to maximize sensitivity.

Chemical structure-based prediction

As an alternative to expression-based classifiers, we used 128 structural features of all chemical compounds as predictors. Random Forest classifiers were repeatedly trained on 70% of the data and tested on the remaining 30%. Classification thresholds were calibrated on the training set to maximize sensitivity.

Genotoxicity

Accuracy %	Specificity %	Sensitivity %	Tissues	Compounds	Accuracy %	Specificity %	Sensitivity %
80.2	92.1	39.8	Liver	110	78.9	95.4	18.6
82.8	90.2	62.5	Cell Culture	69	74.3	90.9	30.5
72.2	85.6	39.5	Kidney	71	74.9	89.3	28.9
80.3	91.3	43.9	All	189	79.6	94.4	26.1

Gene Expression

Carcinogenicity

Chemical Structure

Accuracy %	Specificity %	Sensitivity %	Tissues	Compounds	Accuracy %	Specificity %	Sensitivity %
57.6	37.8	69.1	Liver	110	53.6	22.8	76.4
59.7	53.2	66.0	Cell Culture	69	57.5	19.2	88.6
53.7	56.2	54.0	Kidney	71	66.0	80.3	48.6
58.8	50.6	66.7	All	189	55.5	28.5	78.2

Logistic Regression Model

Genotoxicity

$$\log\left({}^{p(c)}\!\Big/_{1 \quad p(c)} \right) = {}_{0} + {}_{cs}\ \text{Chemical Structure} + {}_{ge}\ \text{Gene Expression}$$

Carcinogenicity

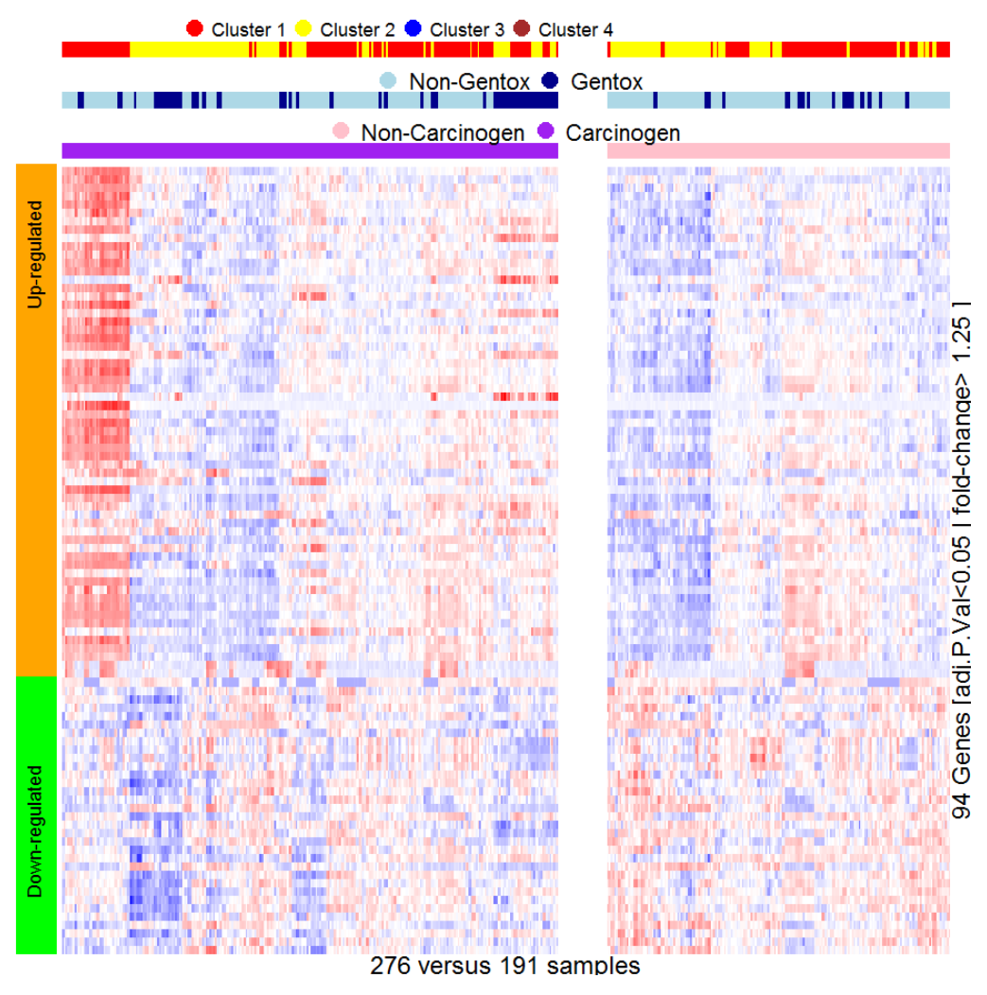
Accuracy %	Specificity %	Sensitivity %	Tissues	Compounds	Accuracy %	Specificity %	Sensitivity %
79.9	91.6	40.1	Liver	110	56.5	37.3	67.7
82.6	90.3	61.7	Cell Culture	69	58.5	50.5	65.3
71.0	85.8	35.4	Kidney	71	57.0	59.4	56.3
80.5	91.0	45.3	All	189	57.8	48.2	66.8

Consensus clustering

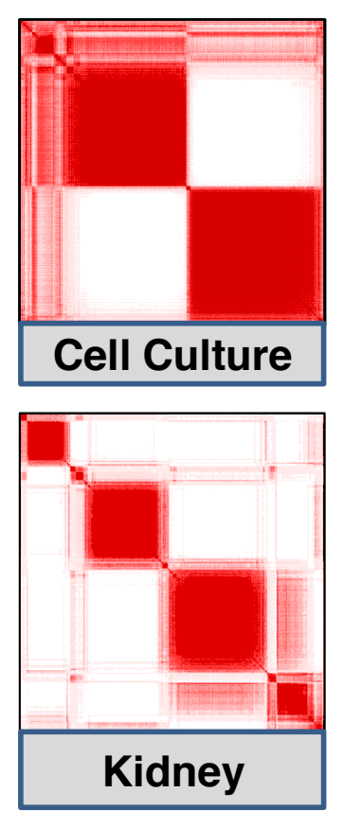
Differential analysis showed the presence of strong substructures within the dataset, reflecting within-class response heterogeneity.

In order to model this substructure, we performed tissue-specific consensus clustering on all of the samples. This analysis identified two reproducible clusters in cell culture, and four clusters in both kidney and liver.

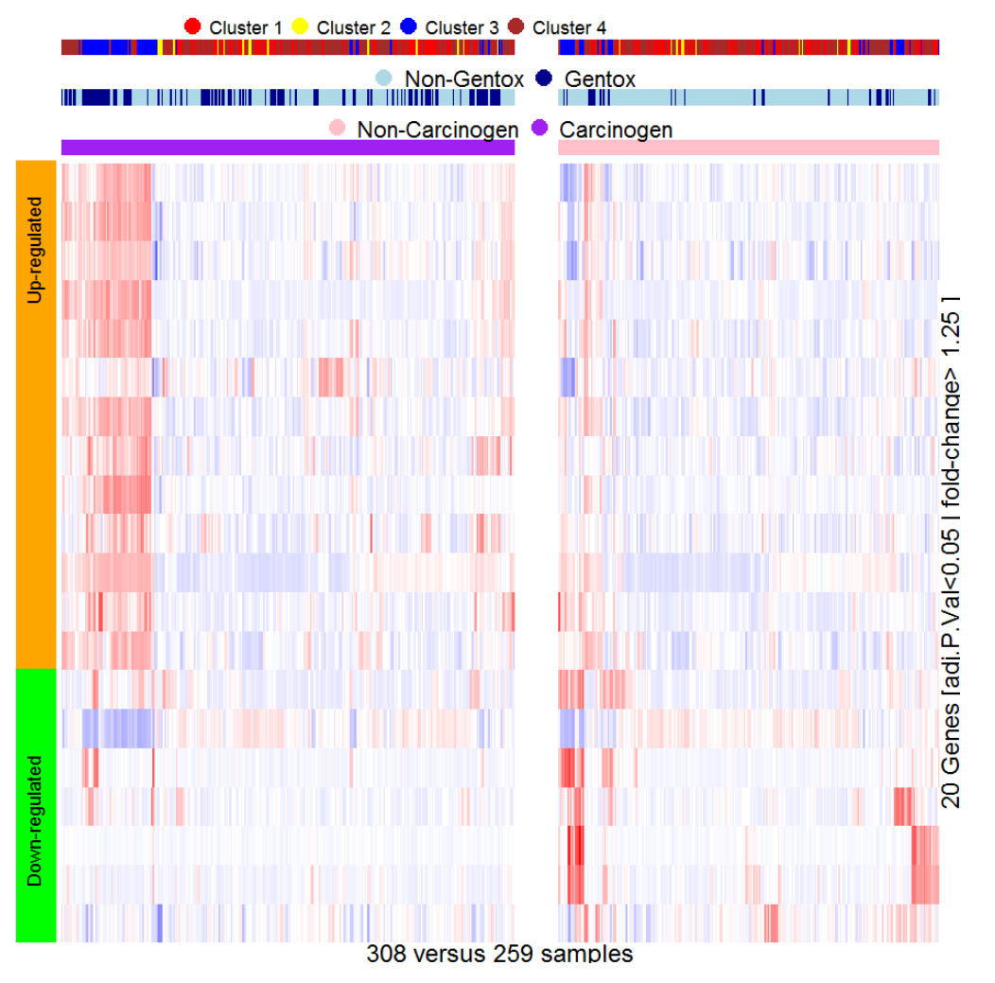
Cell culture: Carcinogenicity



Consensus matrices



Kidney: Carcinogenicity



Conclusions

From this multi-tissue study of gene expression signatures of response to chemical exposure, we observed that:
- Expression response to chemical exposure is tissue specific
- Expression-based and chemical structure-based prediction of *Genotoxicity* is an easier task than predicting *Carcinogenicity*
- Significant within-class response heterogeneity cannot be modeled by a simple binary classifier
- "Sample size" (in terms of # of compounds) is relatively small (~100 or less)

Future Works

In order to improve and extend the methods and preliminary experimental findings obtained thus far, we will:
- Apply classification methodology to larger datasets to incorporate a larger number of compounds during training
- Validate experimental findings on independent datasets such as the (TGgates)
- Incorporate the clustering analysis into our predictive models.

Acknowledgements

This project is a work in progress thanks to the DrugMatrix dataset provided by the National Institute of Environmental Health Science (NIEHS), Scott Auberback of the National Toxicology Program at NIEHS, who annotated a large portion of the chemicals in our dataset, as well as Simone Sciabola from the Bioinformatics Master Program, who provided our group with the structural features for all chemicals.