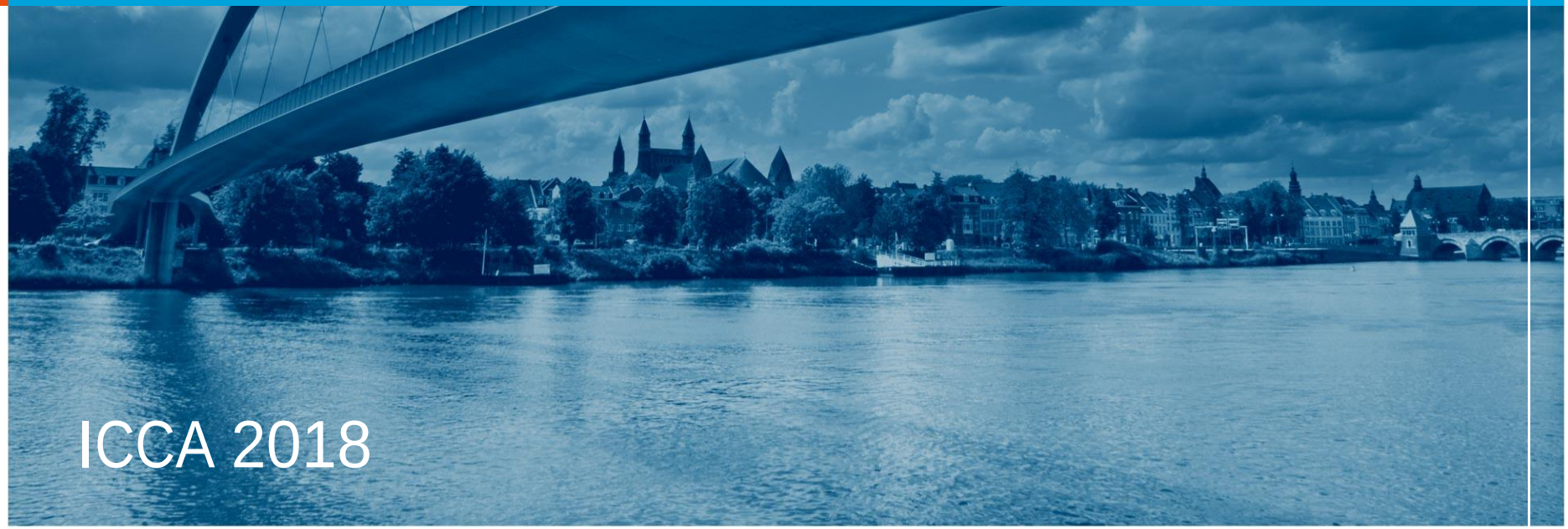
Maastricht University Leading in Learning!

GROW School for Oncology and Developmental Biology

Big Data in Toxicogenomics: Towards FAIR predictions

Danyel Jennen



ICCA 2018

REACH

increasing demands on chemical risk assessment

need for alternative toxicity tests

"Toxicity Testing in the 21th Century"

toxicogenomics

toxicity testing based on high-throughput 'omics' data

alternatives to animal testing

3R's


REACH
need for alternative toxicity tests
Many promising *in vitro* prediction models
alternatives to animal testing
throughput 'omics'
3R's
data

were enough compounds / chemicals used?

What about time points?

is it the right cell model?

or dosages used?

what about the biology of the model?

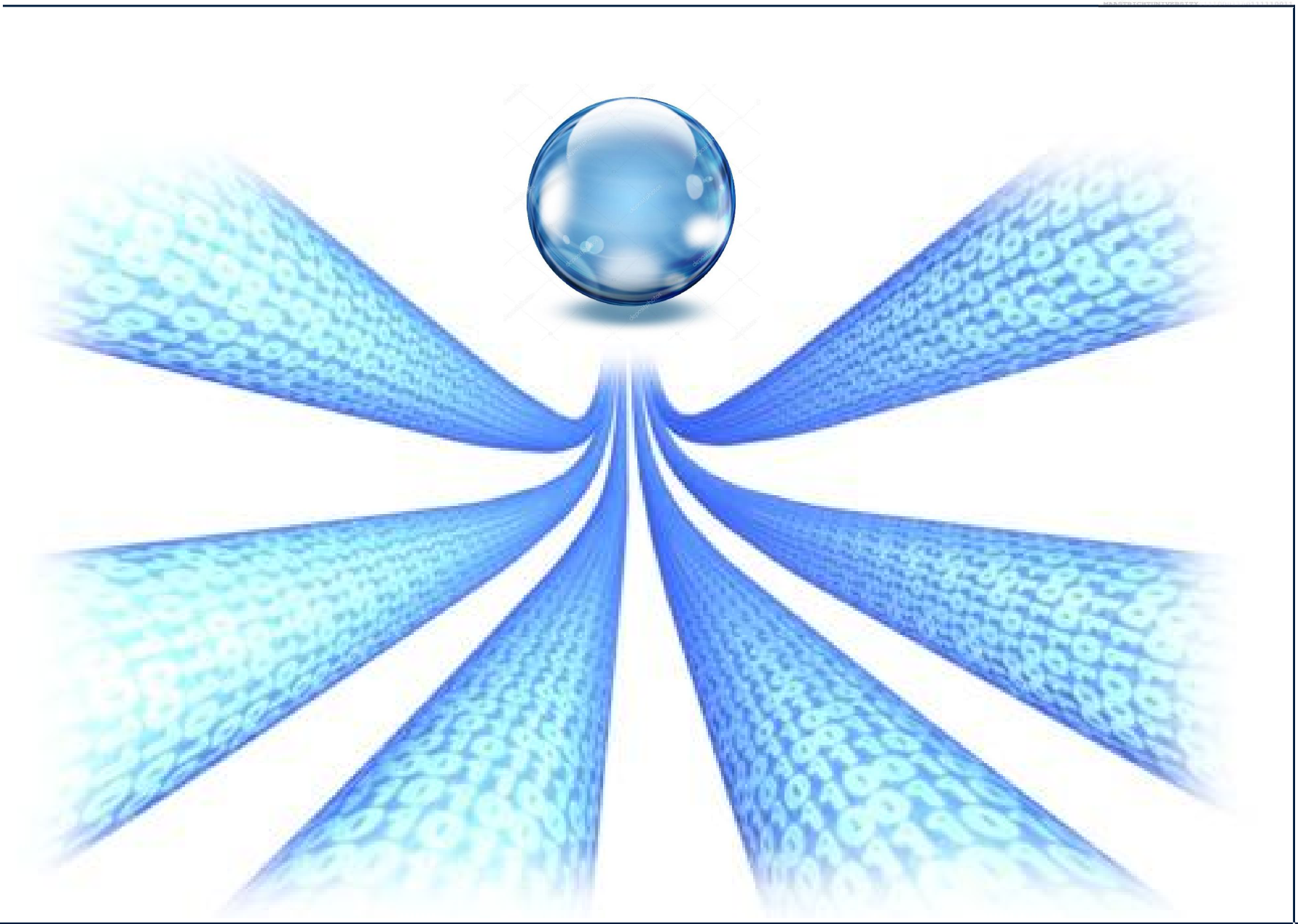
were enough
comp
cher

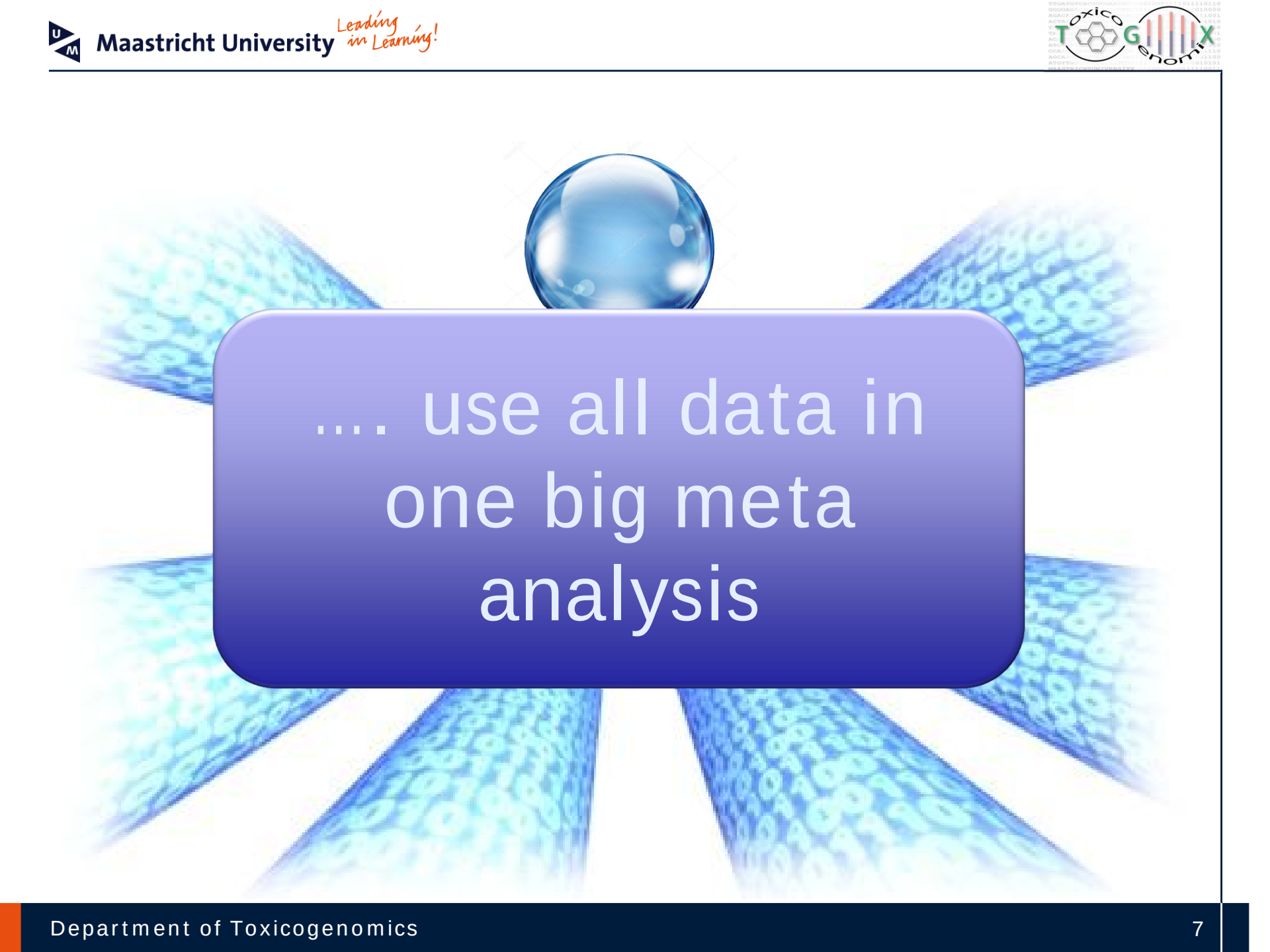
What about time points?

So what if ….

or dosages used?

what about the biology of the model?




.... use all data in one big meta analysis

Aim & explanation of the title

Meta-analysis for *in vivo* genotoxicity prediction using gene expression data from multiple *in vitro* cell models

Big Data In Toxicogenomics:

Towards FAIR predictions

Using gene expression data from multiple toxicity studies stored in freely accessible databases that also provide the corresponding meta-data for a "good" and "honest" prediction

Aim & explanation of the title

Meta-analysis for *in vivo* genotoxicity prediction using gene expression data from multiple *in vitro* cell models

Big Data In Toxicogenomics:

Towards FAIR predictions

Using gene expression data from multiple toxicity studies stored in freely accessible databases that also provide the corresponding meta-data for a "good" and "honest" prediction

Aim & explanation of the title

Meta-analysis for *in vivo* genotoxicity prediction using gene expression data from multiple *in vitro* cell models

Big Data In Toxicogenomics:

Towards FAIR predictions

Using gene expression data from multiple toxicity studies stored in freely accessible databases that also provide the corresponding meta-data for a "good" and "honest" prediction

Aim & explanation of the title

Meta-analysis for *in vivo* genotoxicity prediction using gene expression data from multiple *in vitro* cell models

Big Data In Toxicogenomics:

Towards FAIR predictions

Using gene expression data from multiple toxicity studies stored in freely accessible databases that also provide the corresponding meta-data for a "good" and "honest" prediction

Workflow

Step 1. Data collection

Step 2. Data processing

Step 3. Train prediction model

Step 4. Validate prediction

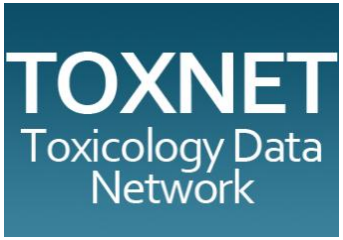
Step 5. Biological interpretation

Step 1. Data collection

Transcriptomics data

Compound information

CEBS



Step 1. Data collection

Source	Cell model	Number of samples (incl. controls)	Number of compounds
diXa data warehouse			
DIXA002 / carcinoGENOMICS	primary rat hepatocytes non-TSA treated	205	15
	primary rat hepatocytes TSA treated	196	15
DIXA-028 / DrugMatrix	primary rat hepatocytes	939	124
Open TG-GATEs	primary rat hepatocytes	3370	145

TSA = trichostatin A

Step 1. Data collection

Source	Cell model	Number of samples (incl.	Number of compounds
diXa da			
DIXA002 carcino			15
			15
DIXA-02			124
Open TG-GATEs	primary rat hepatocytes	3370	145

Total number of unique compounds
235

But, there is only sufficient data available for
In vivo GTX: 24
In vivo NGTX: 45

TSA = trichostatin A

Step 2. Data processing

Affymetrix Rat Genome 230 2.0 Array

RMA normalization
Custom CDF version 22
Ensembl gene IDs

Log2ratios

Averaging biological replicates

619 exposures

205 GTX – 414 NGTX

Split in
training and test sets

80% vs 20%

12162 genes - 619 exposures

10 training/test sets

Step 3. Train prediction model

10 different classification algorithms

R package Caret

Regularized Logistic Regression

Pam: Nearest Shrunken Centroids

Random Forest

k-Nearest Neighbors

Partial Least Squares

GBM: Stochastic Gradient Boosting

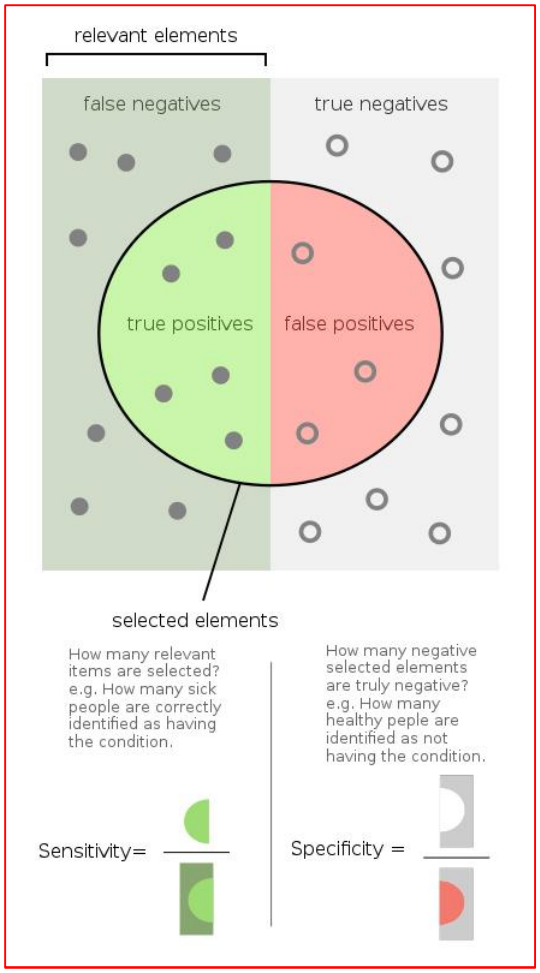
xgbLinear: eXtreme Gradient Boosting

Support vector machines:
svmLinear, svmLinear2, svmLinearWeights

Step 3. Train prediction model

Performance training set run 1

logistic regression and SVM score best



Step 4. Validate prediction

Performance validation set
Average of 10 runs

	PAM	svmL	Logistic
Accuracy	0.776983	0.956457	0.943685
Sensitivity	0.363177	0.910826	0.891524
Specificity	0.968498	0.978004	0.968957

Step 4. Validate prediction

svmL

- Per run between 1 – 12 exposures misclassified
- In total 39 unique exposures misclassified
  - Out of 553 unique exposures
  - Accuracy 93%
- Misclassification usually for extremes
  - Either lowest or highest dosage
  - Shortest or longest exposure
  - No clear relationship between dosage / exposure time and genotoxicity

Step 5. Biological interpretation

svmL

Selection of top genes with at least 60% contribution towards classification

↓

509 genes in at least 1 run
60 genes in all runs

↓

PathVisio
Over-representation analysis

↓

Z-score >1.96
60 genes → 11 significant pathways

Step 5. Biological interpretation

p53 pathway

p53 signal pathway

Relationship between glutathione and NADPH

ATM Signaling Pathway

Genetic alterations of lung cancer

G1 to S cell cycle control

Eicosanoid Synthesis

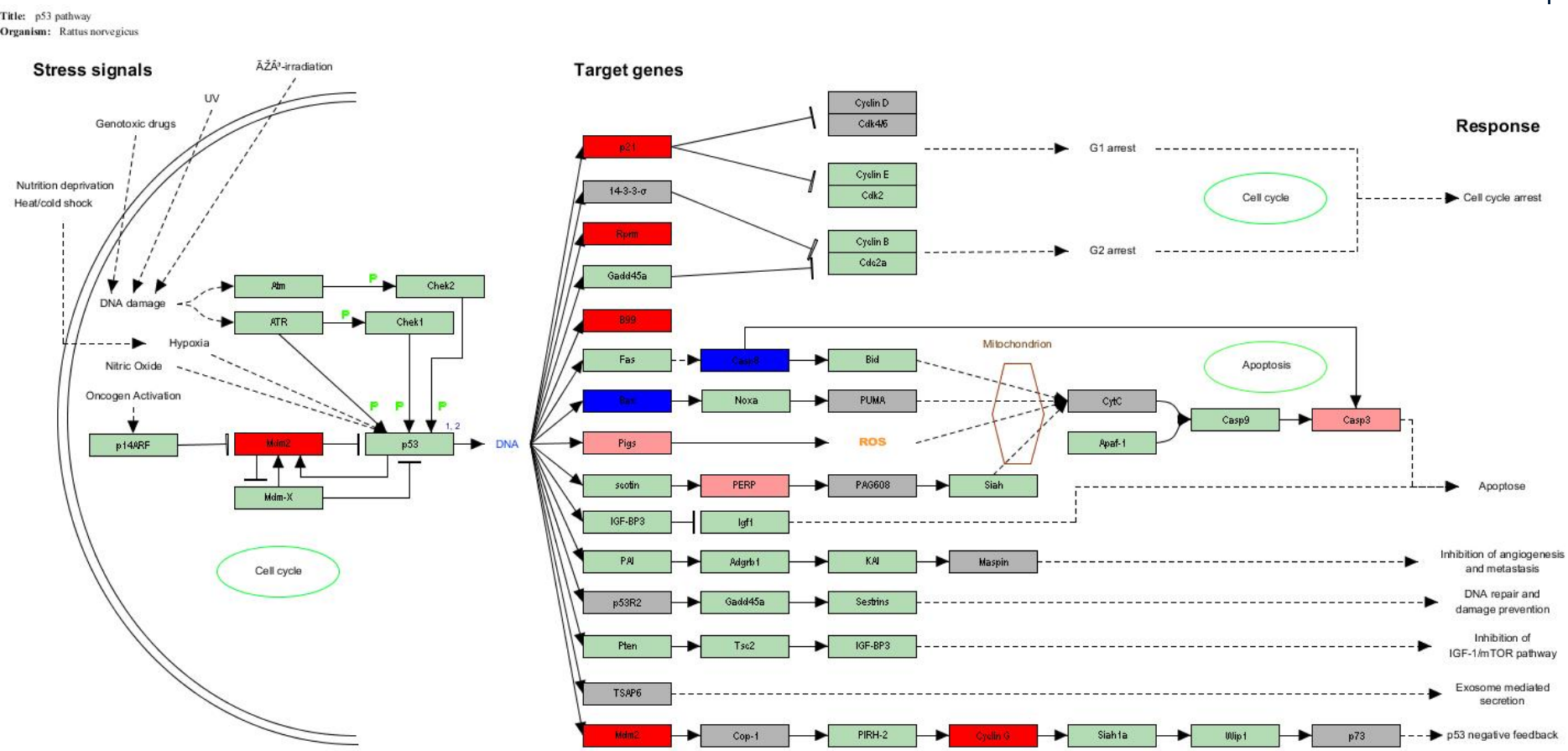
Cell cycle

Fatty Acid Biosynthesis

Selenium Micronutrient Network

Folic Acid Network

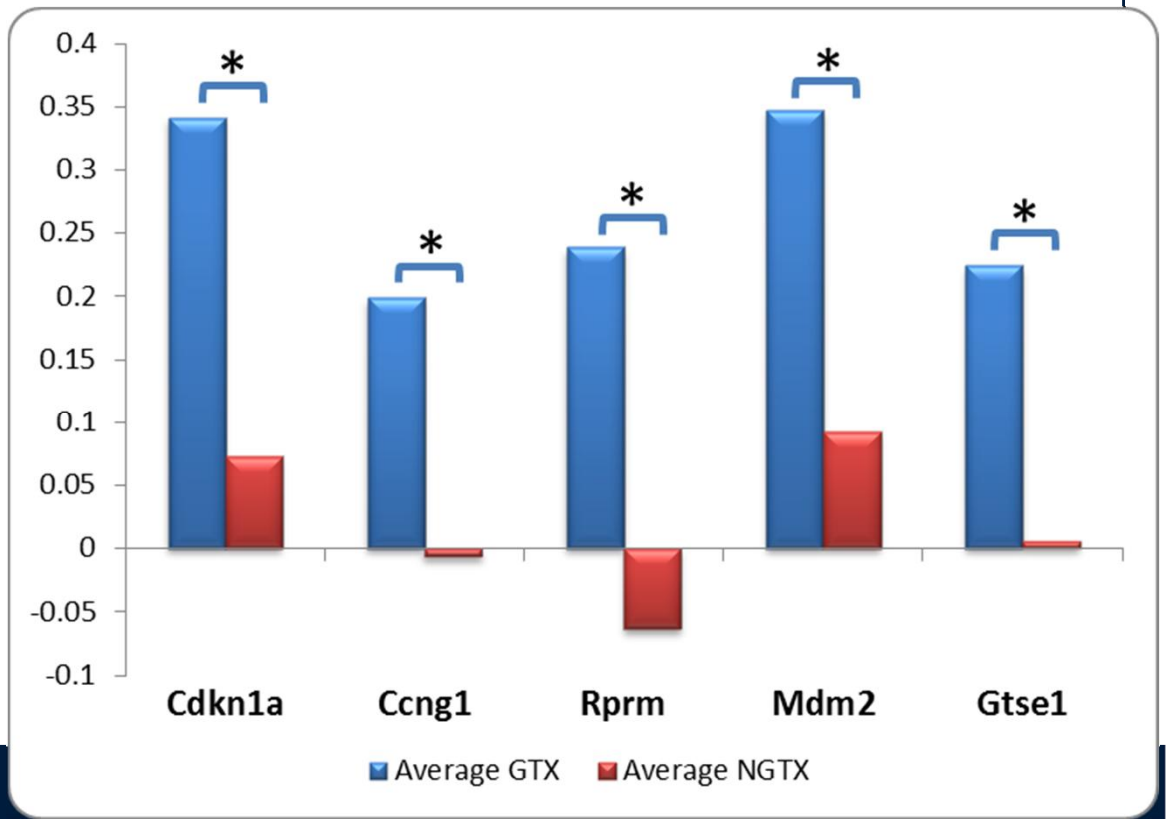
Step 5. Biological interpretation



Present in all 10 runs
Present in 5-9 runs
Present in 1-4 runs
Not present in any run
Not in gene list

Step 5. Biological interpretation

Cdkn1a	cyclin-dependent kinase inhibitor 1A
Ccng1	cyclin G1
Rprm	reprimo, TP53 dependent G2 arrest mediator homolog
Mdm2	MDM2 proto-oncogene
Gtse1	G-2 and S-phase expressed



Summary

- The presented prediction for *in vivo* genotoxicity outperforms the standard *in vitro* test-battery
  - Accuracy ~95% vs ~70%
- In the analysis a fair amount of compounds has been used
  - 69 compounds (24 GTX & 45 NGTX)
- And includes multiple dosages and time-points
  - >600 exposures
- Top features are biologically relevant
  - P53, cell cycle, apoptosis related pathways

Future perspective

- Fine-tune prediction analysis on rat data
  - Add additional compounds / studies
- Proceed with prediction analysis on human data
  - First results presented at IWGT2017
- Proceed with prediction analysis on mouse data
  - Data sets are limited
- Work on automated data and meta-data retrieval
  - Use APIs and containerization (Docker)
- Present final results at Eurotox 2018

Acknowledgement

Juma Bayjanov

Jos Kleinjans

OpenRiskNet

RISK ASSESSMENT E-INFRASTRUCTURE