

# How to benefit from statistics in toxicology and the added value of high-dimensional data

---

Jörg Rahnenführer, Franziska Kappenberg

TU Dortmund University,  
Department of Statistics



Universitätsmedizin  
Göttingen

"Statistical Planning of Translational Studies"  
Symposium

Universität Göttingen, 20.03.-21.03.2024

# Overview: Statistics in toxicology

---

- The role of statistical thinking
- The magical triangle
- Planning of studies
- Analysis pipelines
- Benefit from high-dimensional data (HDD)
- Lots of examples and projects...

# The role of statistical thinking

STATISTICS IN BIOPHARMACEUTICAL RESEARCH  
2023, VOL. 15, NO. 3, 458–467  
<https://doi.org/10.1080/19466315.2023.2224259>



Taylor & Francis  
Taylor & Francis Group



## The Role of Statistical Thinking in Biopharmaceutical Research

Frank Bretz <sup>a,b</sup> and Joel B. Greenhouse <sup>c</sup>

<sup>a</sup>Analytics, Novartis Pharma AG, Basel, Switzerland; <sup>b</sup>Center for Medical Data Science, Institute for Medical Statistics, Medical University of Vienna, Vienna, Austria; <sup>c</sup>Department of Statistics & Data Science, Carnegie Mellon University, Pittsburgh, PA

Quotes on first 6 slides of this talk based on this publication

# The role of statistical thinking



STATISTICS IN BIOPHARMACEUTICAL RESEARCH  
2023, VOL. 15, NO. 3, 458–467  
<https://doi.org/10.1080/19466315.2023.2224259>



Taylor & Francis  
Taylor & Francis Group



## The Role of Statistical Thinking in Biopharmaceutical Research

Frank Bretz <sup>a,b</sup> and Joel B. Greenhouse <sup>c</sup>

<sup>a</sup>Analytics, Novartis Pharma AG, Basel, Switzerland; <sup>b</sup>Center for Medical Data Science, Institute for Medical Statistics, Medical University of Vienna, Vienna, Austria; <sup>c</sup>Department of Statistics & Data Science, Carnegie Mellon University, Pittsburgh, PA

Quotes on first 6 slides of this talk based on this publication

In biopharmaceutical research, toxicology is often the first step, and “the practice of statistics is built on the foundation of good statistical thinking”.

# The role of statistical thinking

---

- Four general steps of problem-solving:
  - A) Understanding and representing the problem
  - B) Determining the data strategy, including an inventory of available data and possibly the collection of additional data
  - C) Developing and executing a solution strategy, including exploratory analysis and model building
  - D) Evaluating and communicating the research results
- Bretz and Greenhouse show how these steps align well with the way “clinical biostatisticians typically engage in collaborative clinical research”

# The role of statistical thinking

STATISTICS IN BIOPHARMACEUTICAL RESEARCH  
2023, VOL. 15, NO. 3, 458–467  
<https://doi.org/10.1080/19466315.2023.2224259>



Taylor & Francis  
Taylor & Francis Group



## The Role of Statistical Thinking in Biopharmaceutical Research

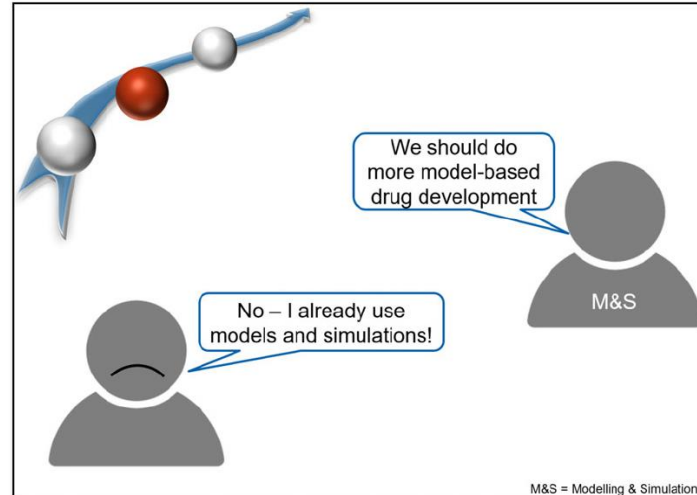
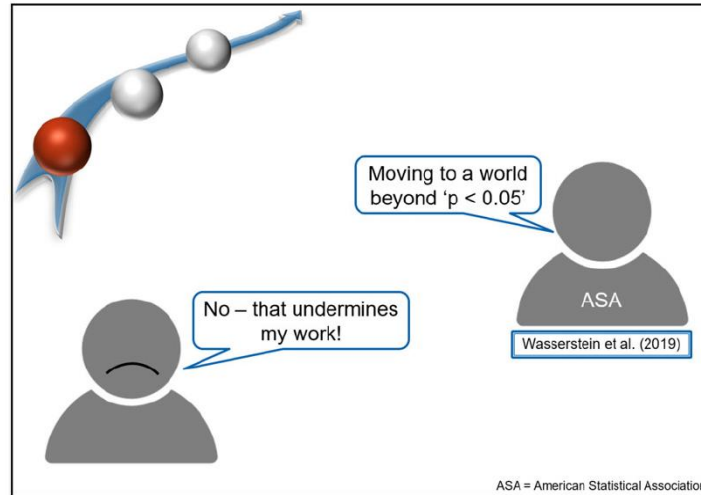
Frank Bretz <sup>a,b</sup> and Joel B. Greenhouse <sup>c</sup>

<sup>a</sup>Analytics, Novartis Pharma AG, Basel, Switzerland; <sup>b</sup>Center for Medical Data Science, Institute for Medical Statistics, Medical University of Vienna, Vienna, Austria; <sup>c</sup>Department of Statistics & Data Science, Carnegie Mellon University, Pittsburgh, PA

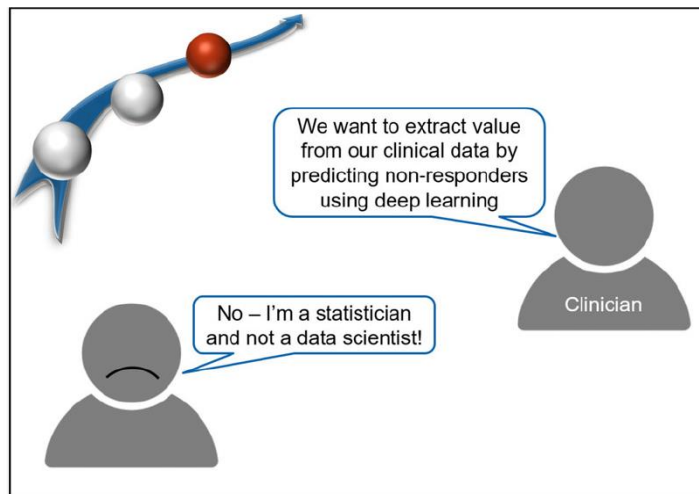
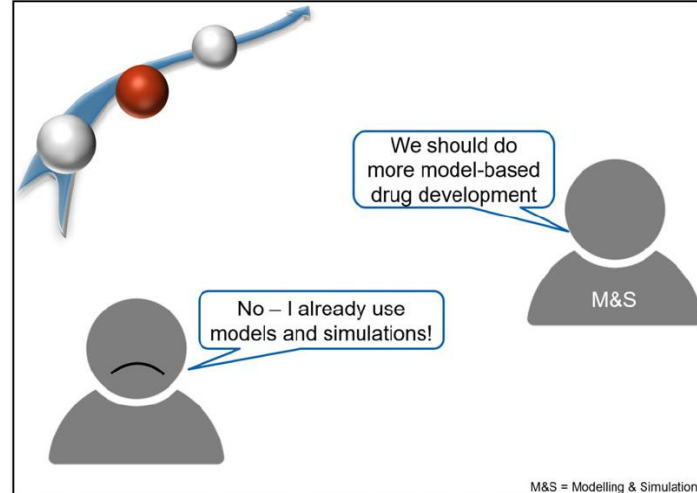
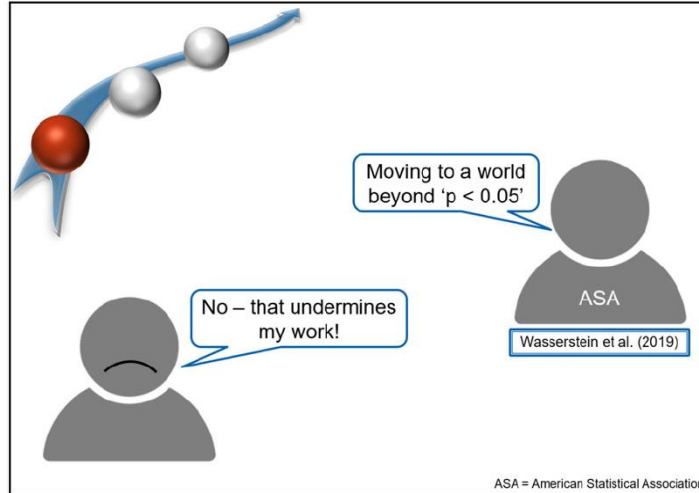
- The ideas outlined in this article are not new. For example, Box (1976) wrote:

successful collaboration requires the wit to comprehend complicated scientific problems, the patience to listen, the penetration to ask the right questions, and the wisdom to see what is, and what is not, important.

# The role of statistical thinking



# The role of statistical thinking



Cycles of innovation, including the current emergence of big data and machine learning, offer new opportunities for statisticians!

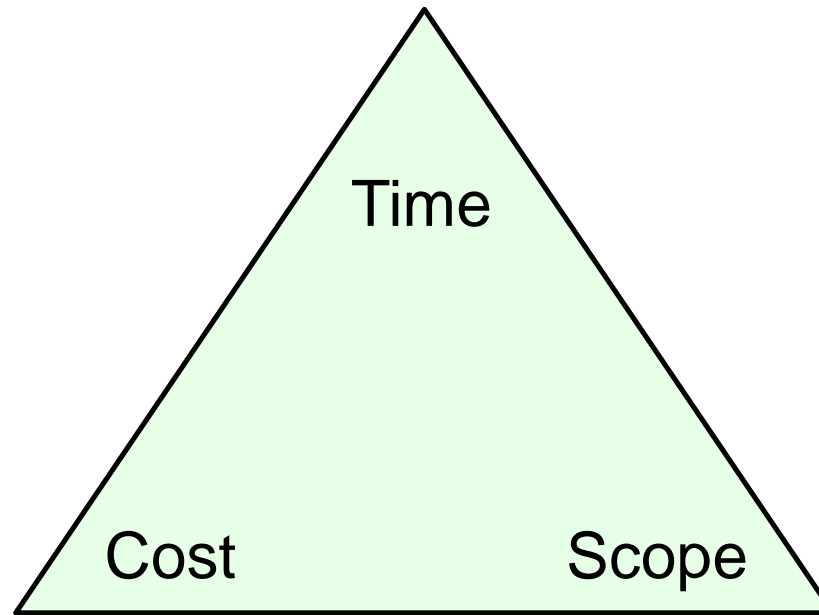
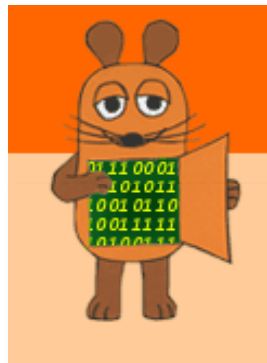


# The magical triangle



Informatics

Statistics

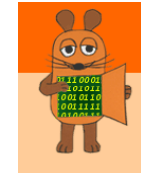
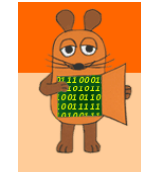


Toxicology

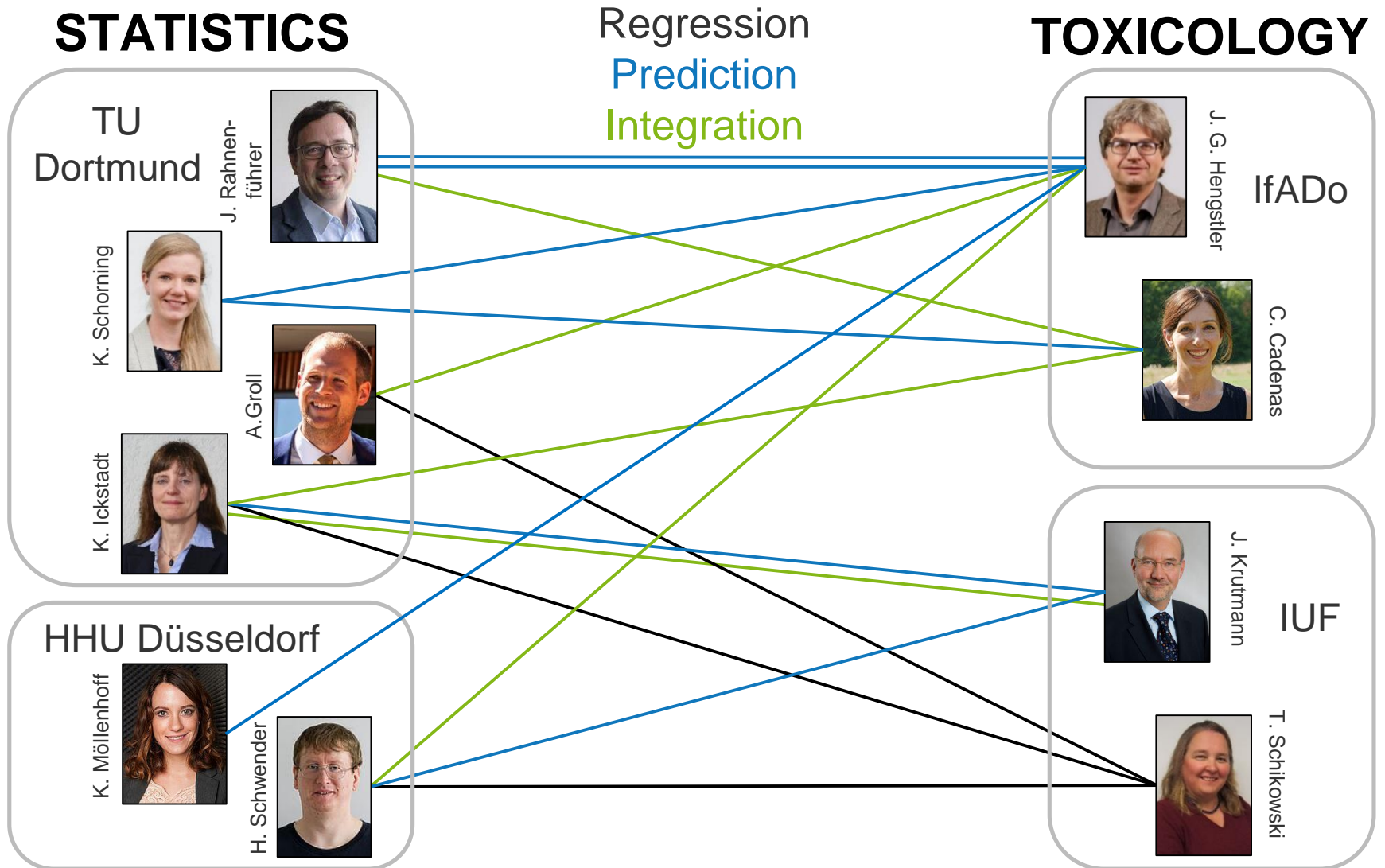


# Data analysis for experimental studies

1. Toxicological question
2. Experimental design
3. Toxicological experiment
4. Data preprocessing
5. Statistical analysis
6. Biological verification and interpretation



# RTG 2624 – Interdisciplinarity



# Interdisciplinarity – statistical thinking

---

- Now more concrete and some examples
- Three short stories
  - “Mixing up” samples
  - “Omitting” bad experiments
  - “Forgetting” controls

# Interdisciplinarity – statistical thinking

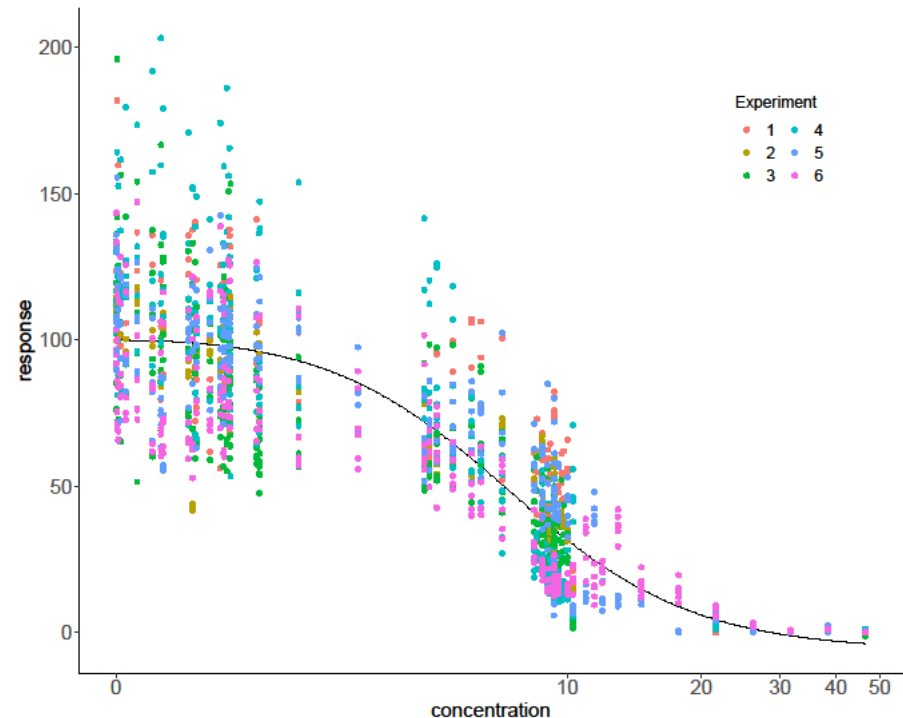
---

- Now more concrete and some examples
- Three short stories
  - “Mixing up” samples
  - “Omitting” bad experiments
  - “Forgetting” controls

- Our statistical contributions
  - Design
  - Guidance
  - Quality control
  - Analysis pipelines
  - New analysis approaches with HDD

# Statistical thinking: Design

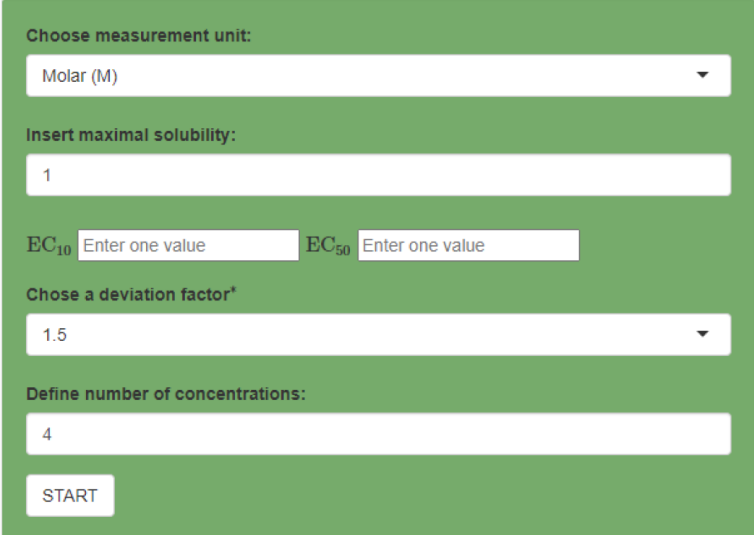
- Design of optimal concentrations for in vitro cytotoxicity experiments (Leonie Schürmeyer, ..., Jan G. Hengstler, Kirsten Schorning)
- Goal: Modelling of sigmoidal dose-response curve
- Statistical criteria for choosing concentration (D-optimal, Bayesian D-optimal, etc.)
- Experiments were performed by toxicologists for large number of concentrations, to support statistical method development



# Statistical thinking: Design

- **Design of optimal concentrations** for in vitro cytotoxicity experiments (Leonie Schürmeyer, ..., Jan G. Hengstler, Kirsten Schorning)
- Goal: Modelling of sigmoidal dose-response curve
- Statistical criteria for choosing concentration (D-optimal, Bayesian D-optimal, etc.)
- **Experiments were performed by toxicologists for large number of concentrations, to support statistical method development**

- Guidance: R Shiny app: **Optimal concentrations for cytotoxicity experiments**



The screenshot shows a web-based interface with a green background. It contains several input fields and a button:

- Choose measurement unit:** A dropdown menu with "Molar (M)" selected.
- Insert maximal solubility:** A text input field containing the value "1".
- EC<sub>10</sub>:** A text input field with the placeholder "Enter one value".
- EC<sub>50</sub>:** A text input field with the placeholder "Enter one value".
- Chose a deviation factor\*:** A dropdown menu with "1.5" selected.
- Define number of concentrations:** A text input field containing the value "4".
- START:** A button located at the bottom left of the form.

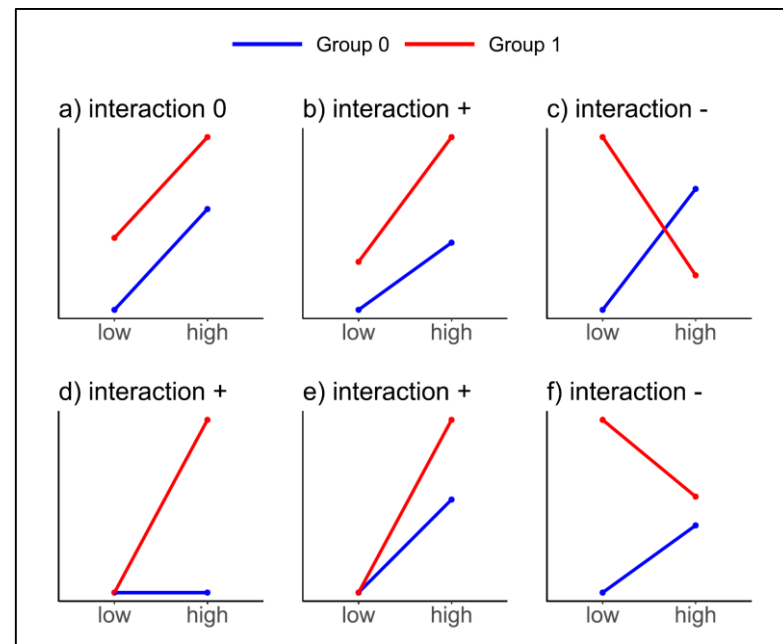
- EC values are pre-specified, and for robustification a deviation factor can be allowed

# Statistical thinking: Guidance II

- Benefit of using interaction effects for the analysis of high-dimensional time-response or dose-response data for two-group comparisons

Julia C. Duda, Carolin Drenda, Hue Kästel, Jörg Rahnenführer, Franziska Kappenberg. *Scientific Reports* 13(1): 20804, 2023

- Figure: Schematic depiction of data scenarios without and with interaction
- Common “mistake”: Check if “no significant effect” is observed for low (dose/time) and “significant effect” is observed for high
- Better: directly calculate significance of interaction effect!





# Statistical thinking: Guidance III

---

- **Guidance for statistical design and analysis of toxicological dose-response experiments, based on a comprehensive literature review**

Franziska Kappenberg, Julia C. Duda, Leonie Schürmeyer, Onur Gül, Tim Brecklinghaus, Jan G. Hengstler, Kirsten Schorning, Jörg Rahnenführer.  
*Archives of Toxicology* 97, 2741-2761, 2023

---

<b>D</b>	Design	Plan the design of the experiment, considering the analysis plan
<b>E</b>	Experiment	Conduct the toxicological experiment
<b>N</b>	Normalize	Perform normalization, tailored to the type of assay (e.g., remove batch effects, convert to percentages, etc.)
<b>M</b>	Modelling	Model the dose–response relationship; if possible consider fitting a parametric model
<b>A</b>	Alert concentration	Calculate the alert concentration of interest (e.g., ED values, NOAEL, BMD, LOEC (Dunnett-type test))
<b>R</b>	Report	Report precisely all applied methods (testing/modelling) and the resulting conclusions

---

# Statistical thinking: Guidance IV

---

- **MoS-TEC: A toxicogenomics database based on model selection for time-expression curves**

Franziska Kappenberg, Benedikt Kütke, Jörg Rahnenführer,  
*submitted, 2024*

- TG-GATEs (Open Toxicogenomics Project-Genomics Assisted Toxicity Evaluation System) database provides genomewide gene expression data for 170 compounds
- **Time-response analysis possible for each gene separately for 8 time points and 4 doses (including control)**
- Challenge: Automation for thousands of genes
- Crucial step: Preprocessing and quality control steps

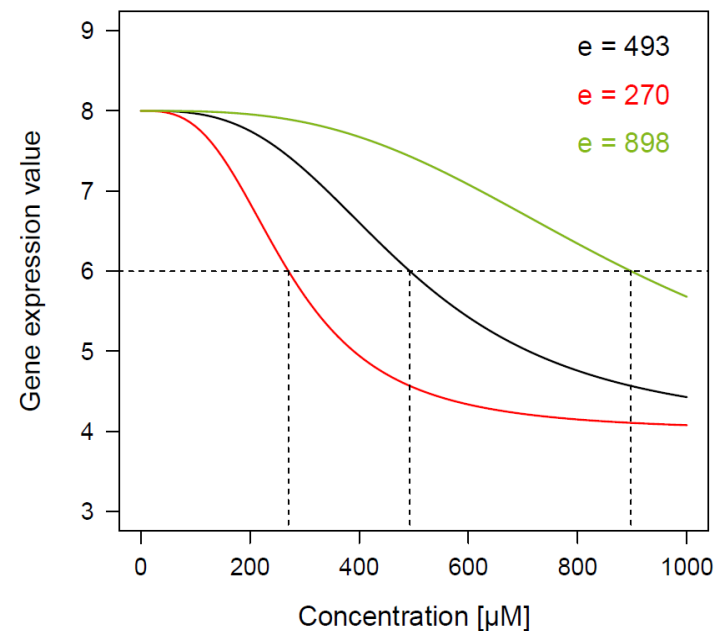
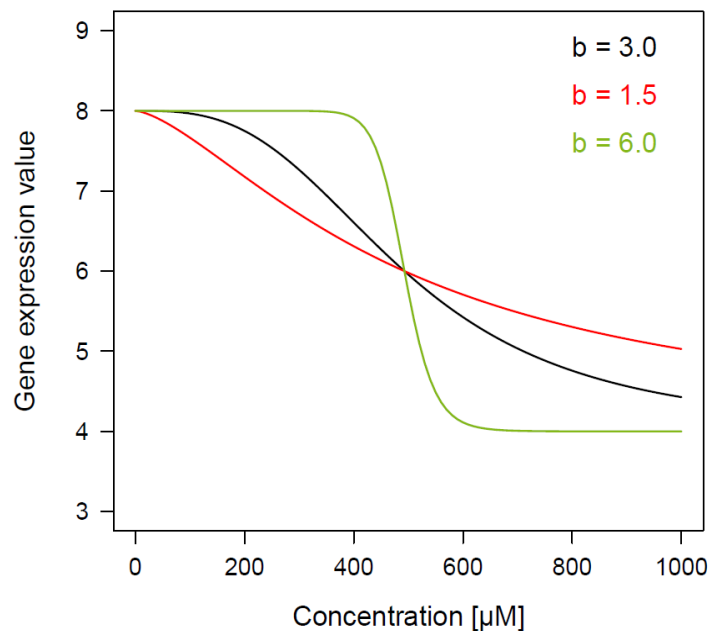
# Statistical thinking: Dose-response analysis

## The four-parameter log-logistic model (4pLL model)

- Let  $x > 0$  be a concentration and  $\phi = (b, c, d, e)$  a parameter vector with  $e > 0$ . The 4pLL model is given by:

$$f_{4pLL}(x, \phi) = c + \frac{d - c}{1 + \exp(b(\log(x) - \log(e)))}$$

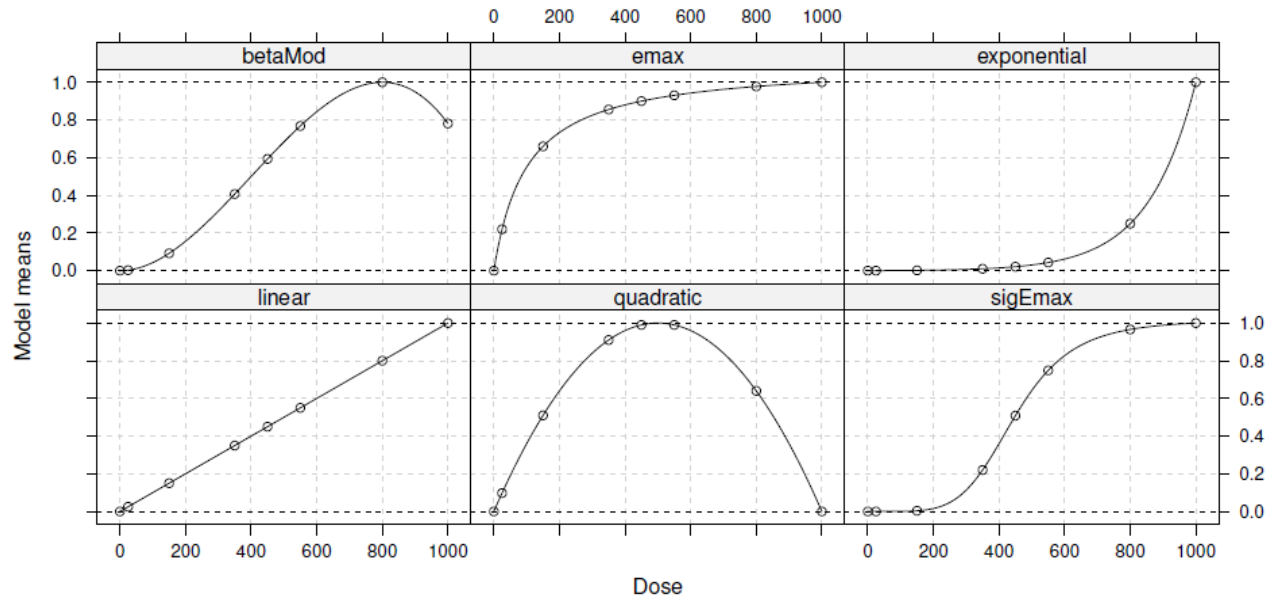
- Parameter  $e$  : ED50 parameter, i.e. corresponding to half-maximal effect



# Statistical thinking: Dose-response analysis

- **Model selection**

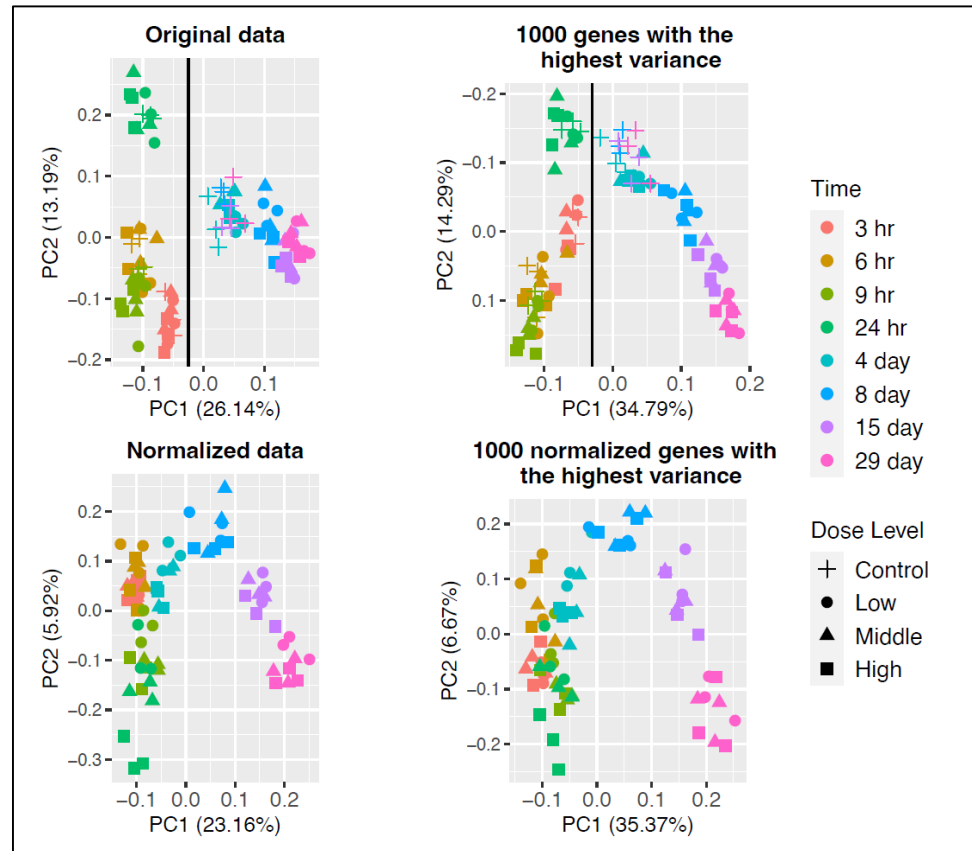
- Identify best model per gene with MCP-Mod (multiple comparison procedure + modelling), two-step procedure originally developed for Phase II clinical studies
- ‘Guesstimates’ for candidate models are required for all genes simultaneously



# Statistical thinking: Time-response analysis

## • Preprocessing

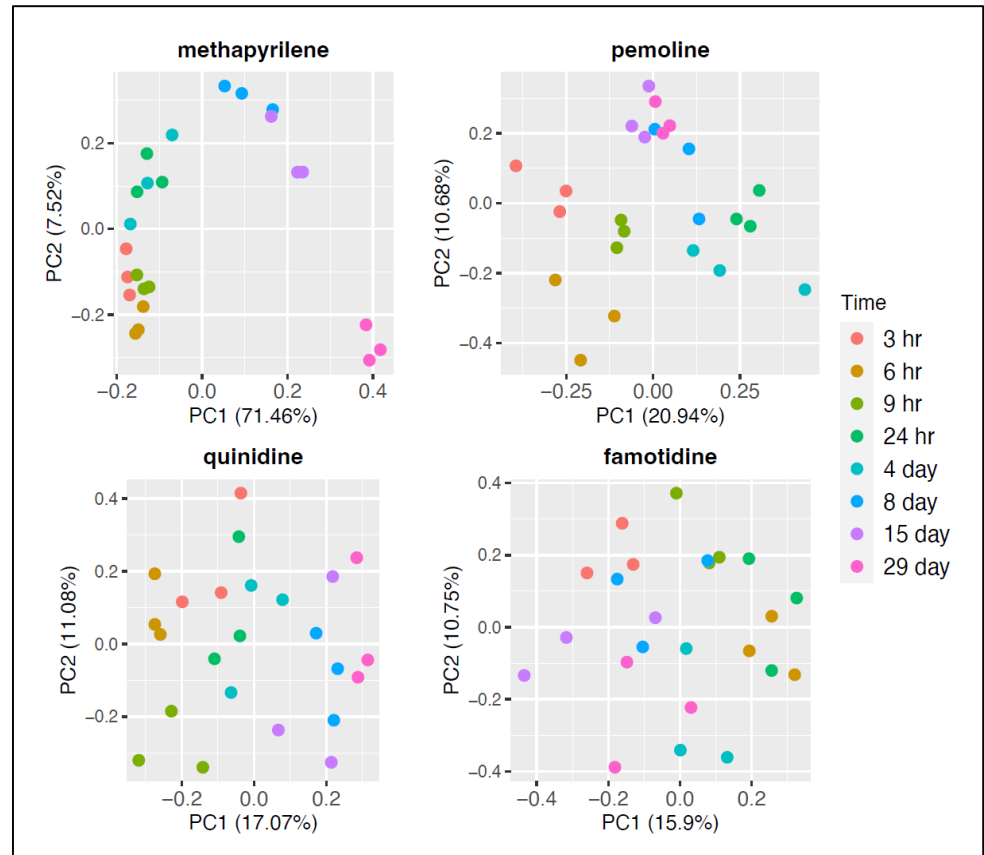
- PCA plots for compound propylthiouracil, all dose levels, all time points
- Original vs. normalized data (top, bottom), with and without gene selection (left, right)
- Normalization and gene selection reduce batch effect and noise



# Statistical thinking: Time-response analysis

## • Preprocessing

- Compound 1: methapyrilene:  
Clear progression of time points
- Compound 2: pemoline:  
Some progression along time points
- Compound 3: quinidine:  
Random noise overlays structure
- Compound 4: famotidine:  
Almost no structure



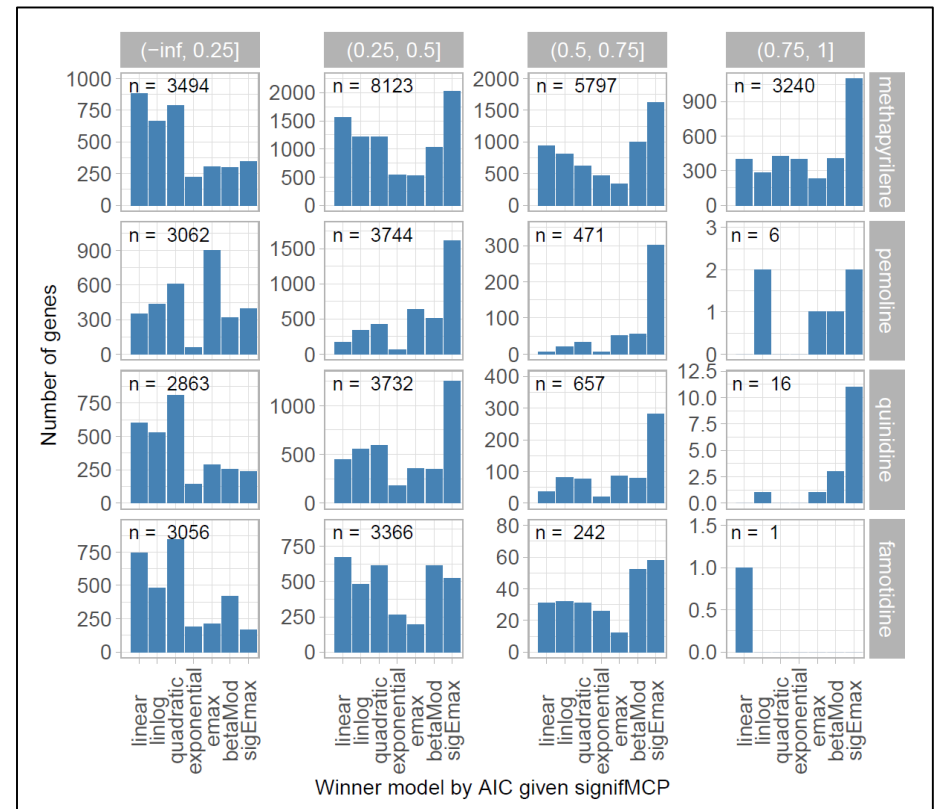
# Statistical thinking: Time-response analysis

- **Model selection**

- Identify best time-response model per gene with MCP-Mod (multiple comparison procedure + modelling)
- Rows: Compounds  
Columns: Adjusted  $R^2$  binning
- More significant genes for compounds with clear structure in PCA

- **Insight**

- Best model depends on gene/compound combination, sigmoidal model often suitable for high-quality fits



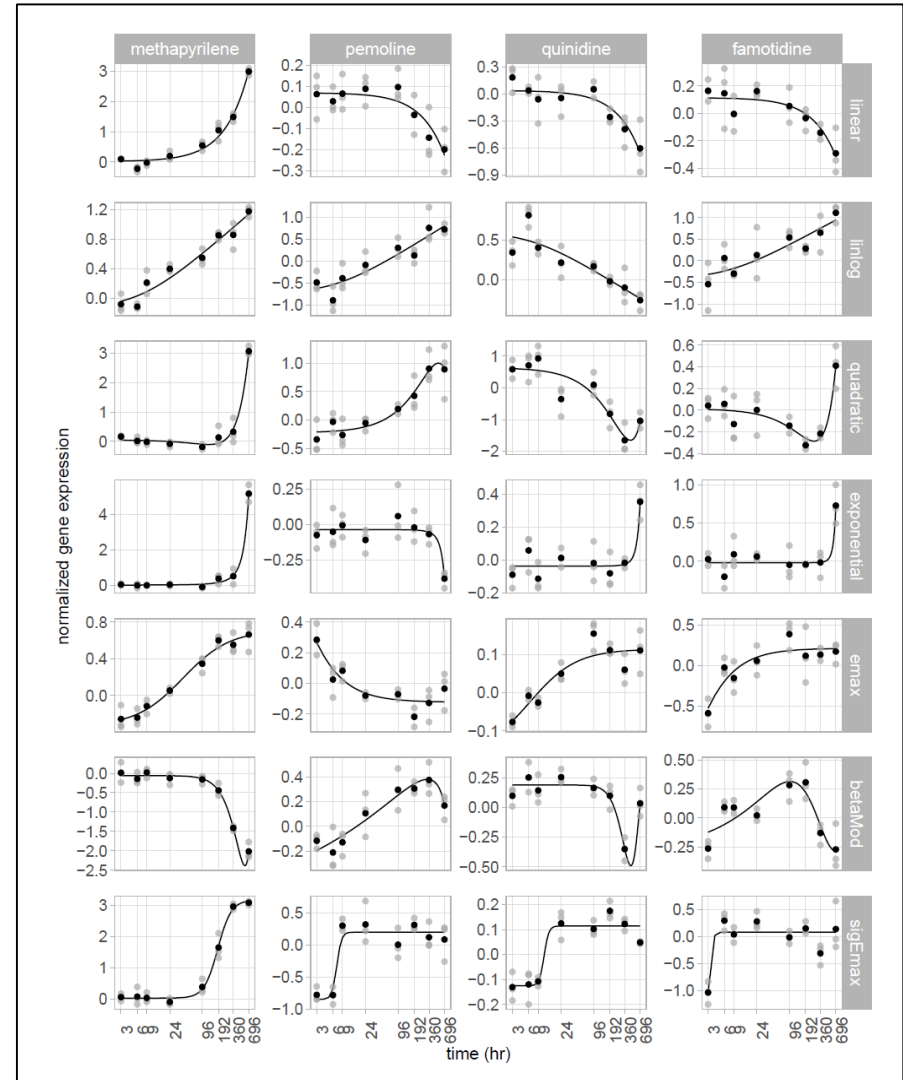
# Statistical thinking: Time-response analysis

- **Model selection**

- Identify best time-response model per gene with MCP-Mod (multiple comparison procedure + modelling)
- Rows: Compounds  
Columns: Adjusted  $R^2$  binning
- More significant genes for compounds with clear structure in PCA

- **Example genes**

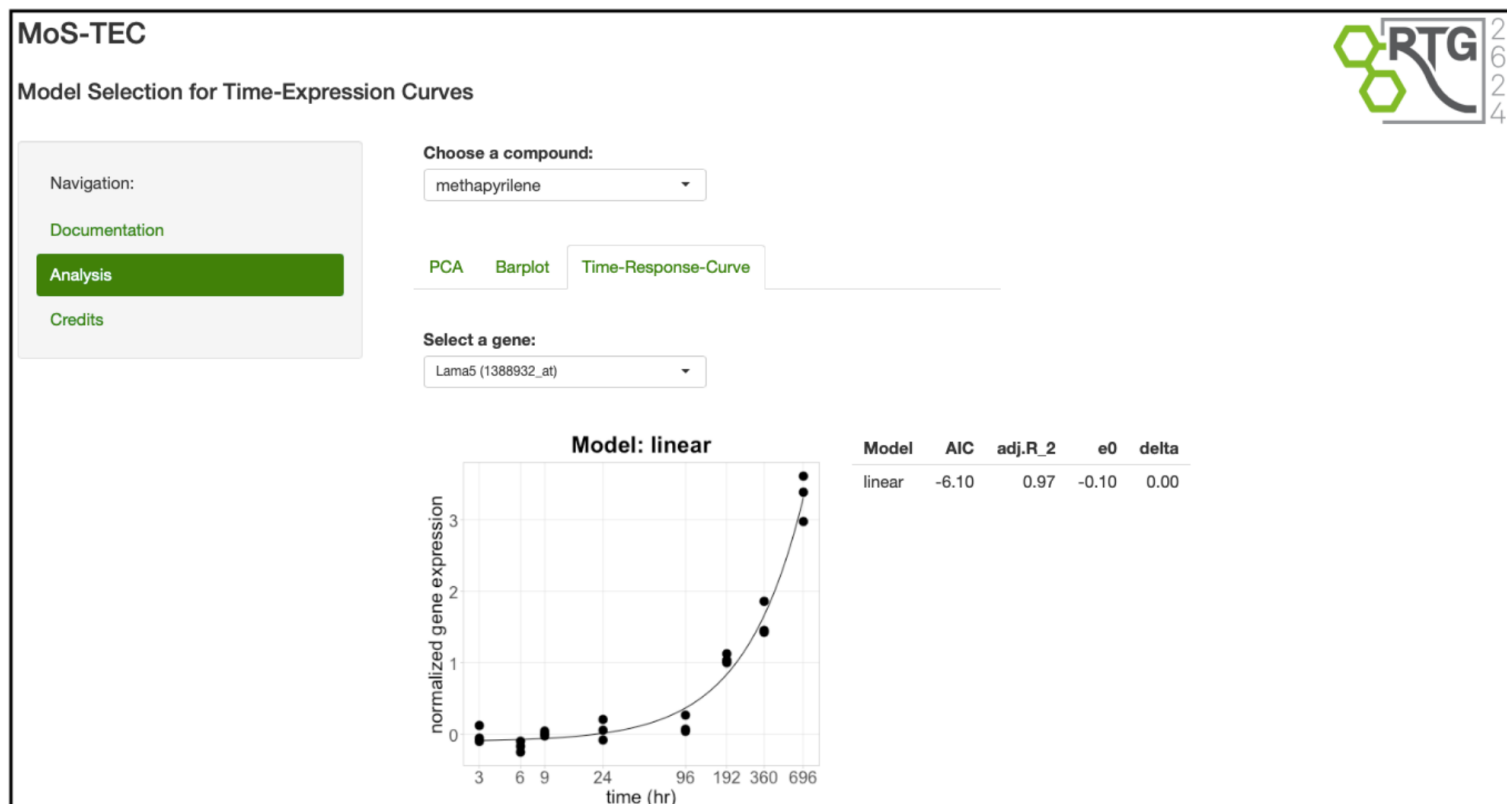
- Examples per compound and model, where the respective model yields the best fit – *logarithmic scale on x-axis!*





# Statistical thinking: Time-response analysis

- R Shiny app: [MoS-TEC: Model Selection for Time-Expression Curves](http://shiny.statistik.tu-dortmund.de:8080/app/MoS-TEC)  
<http://shiny.statistik.tu-dortmund.de:8080/app/MoS-TEC>



# Statistical thinking: Analysis pipeline

---

- Spatio-Temporal Multiscale Analysis of Western Diet-Fed Mice Reveals a Translationally Relevant Sequence of Events during NAFLD Progression

Ahmed Ghallab, ..., Julia Duda, ..., Franziska Kappenberg, ..., Jörg Rahnenführer, ..., Jan G Hengstler. *Cells* 10, 2516, 2021

- Scientific goal: Understand non-alcoholic fatty liver disease (NAFLD)
- Comparison of two groups of mice (mouse model)
  - SD (standard diet) and WD (Western diet, mimics fast-food-style diet)
  - Measurements after week 3, 6, 12, 18, 24, 30, 36, 42, 48

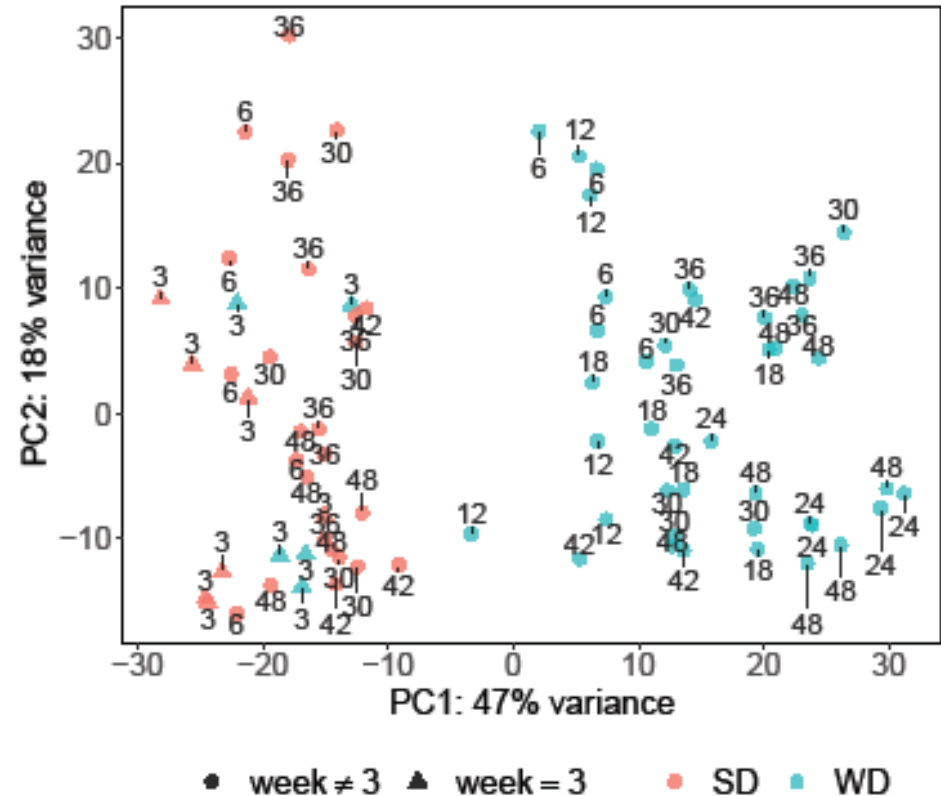
# Statistical thinking: Analysis pipeline

- **Visualization**

- PCA plot of RNAseq data (top 500 most variable genes)
- WD (blue) and SD (red) mice
- Numbers indicate weeks

- **Insight**

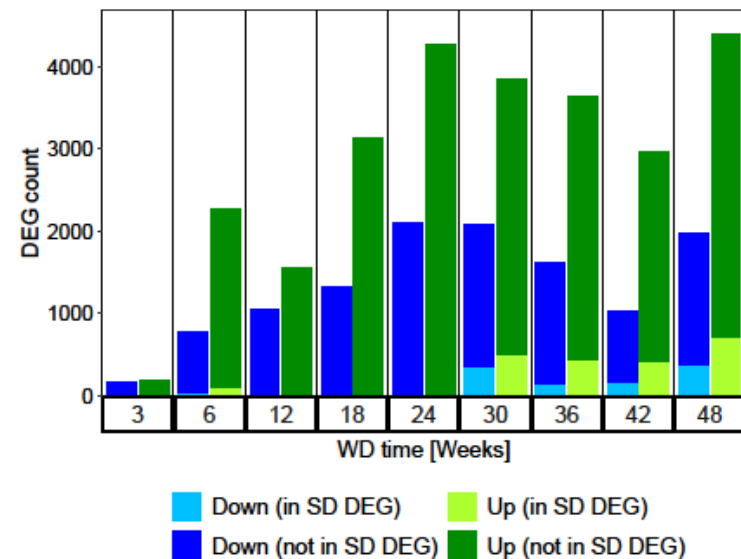
- Data for WD mice similar after 3 weeks, but very different from 6 weeks on



# Statistical thinking: Analysis pipeline

- Calculation of **differentially expressed genes**

- Comparison always against “SD, 3 weeks” (control)
- DEG count: numbers of differentially expressed genes
  - Adjusted p-value and fold change considered, i.e. biological relevance and statistical significance
- Light blue and light green: DEGs in SD (and WD) for time periods with SD controls
- Dark blue and dark green: DEGs in WD but not in SD



- Result: Many additional genes for WD
- **Subsequently: Analyze interactions effects between diet and time point**

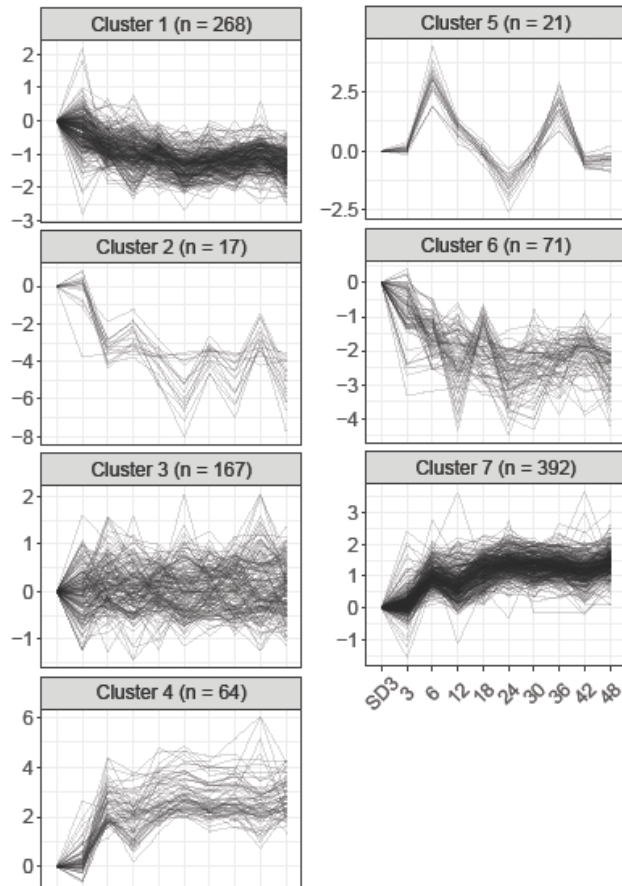
# Statistical thinking: Analysis pipeline

---

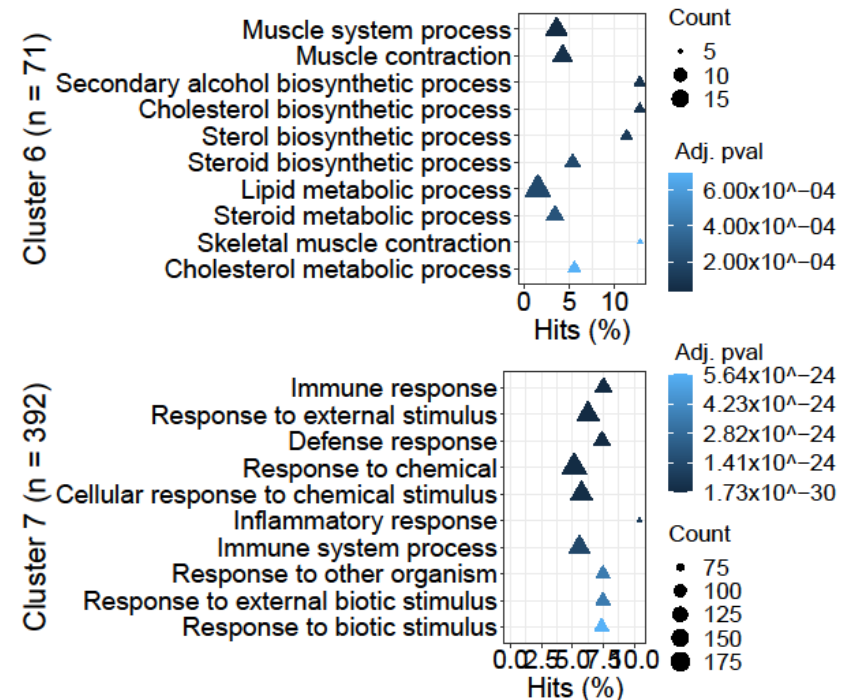
- Goal: Understand mechanisms on genetic level
  - No insight into relevance or role of single genes
  - Strong technical and biological noise and curse of dimensionality
- Improvement: Adding other biological information
- **Gene group tests**
  - Idea: **Grouping of genes that have a known, usually functional relationship**
  - Examples: **Gene Ontology groups**, KEGG pathways, groups of interesting genes from previous studies
  - Start with ordered gene list (here: differential expression), then two statistical approaches for scoring a group G:
    - Count number of members of G below and above prefixed cutpoint
    - Analyze distribution of ranks of members of group G with KS test

# Statistical thinking: Analysis pipeline

- K-means clustering of expression data

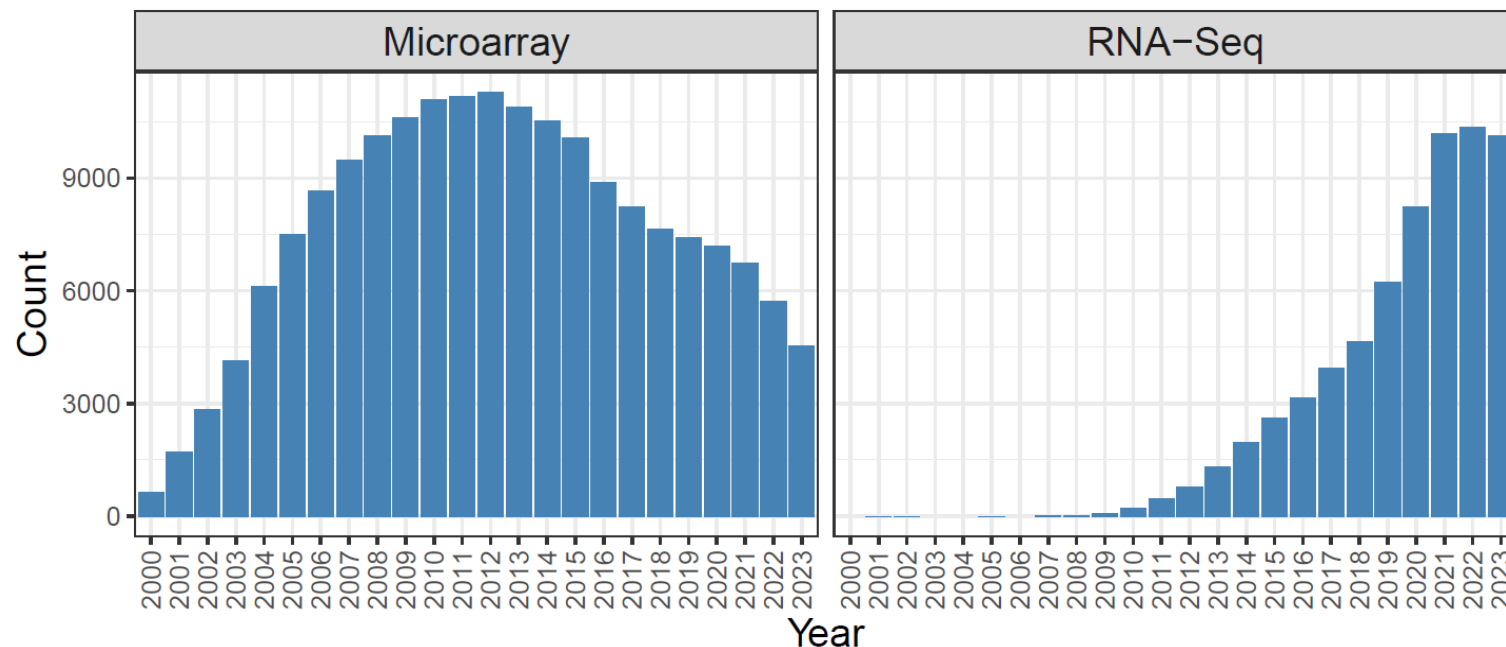


- Analysis of most enriched GO groups per cluster
- Toxicological interpretation important



# Availability of high-dimensional data

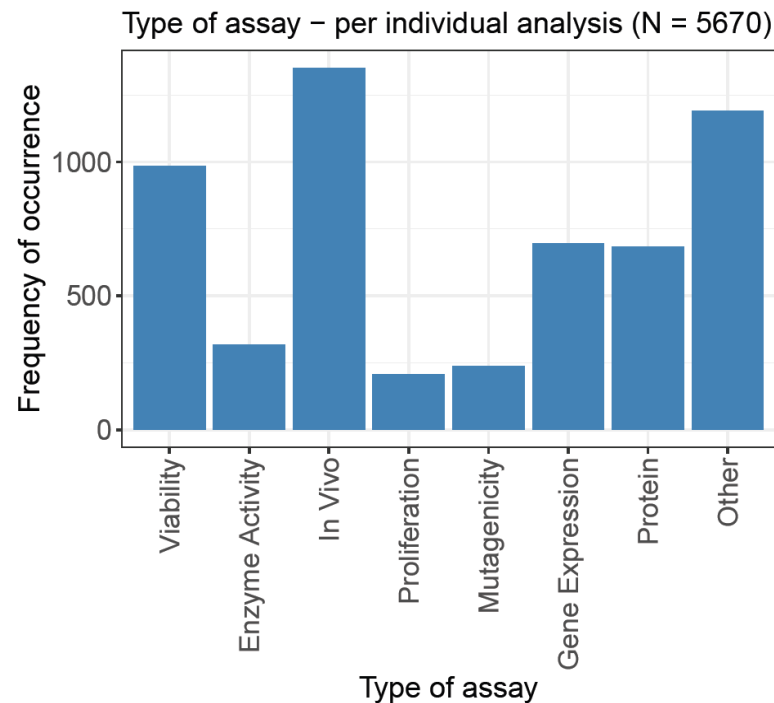
- Relevance of technologies for measuring high-dimensional gene expression data
- Number of publications over time: PubMed search for the keywords 'Microarray' and 'RNA-Seq', January 2024



# Availability of high-dimensional data

- Guidance for statistical design and analysis of toxicological dose-response experiments, based on a comprehensive literature review  
Franziska Kappenberg et al., 2023

- Review of all dose-response curves published in three major toxicological journals in 2021
- Gene expression and protein data often available, but mainly analyzed via barplots

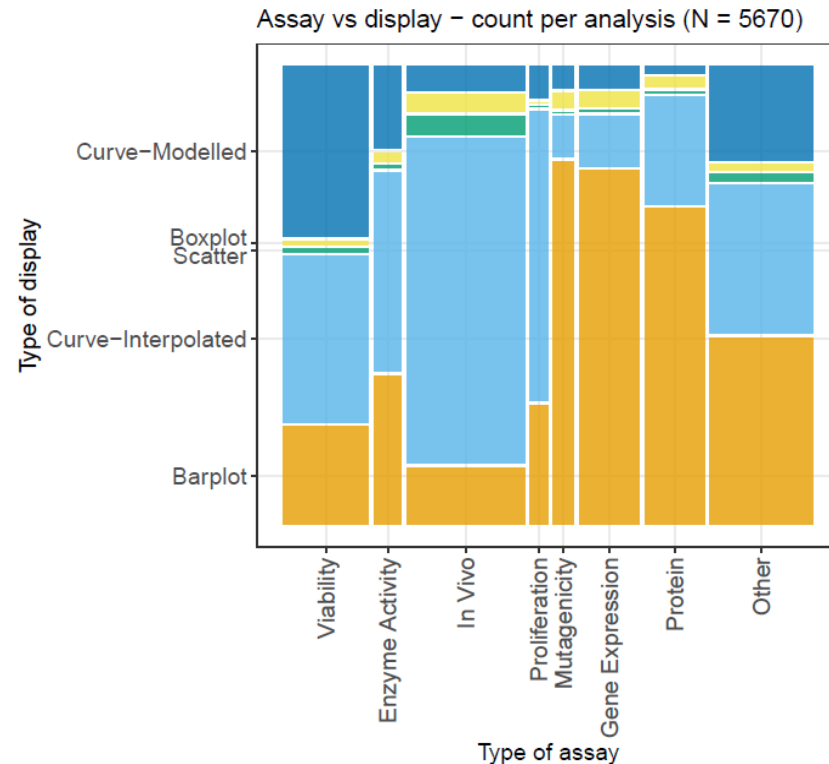




# Availability of high-dimensional data

- **Guidance for statistical design and analysis of toxicological dose-response experiments, based on a comprehensive literature review**  
Franziska Kappenberg et al., 2023

- Review of all dose-response curves published in three major toxicological journals in 2021
- Gene expression and protein data often available, but mainly analyzed via barplots

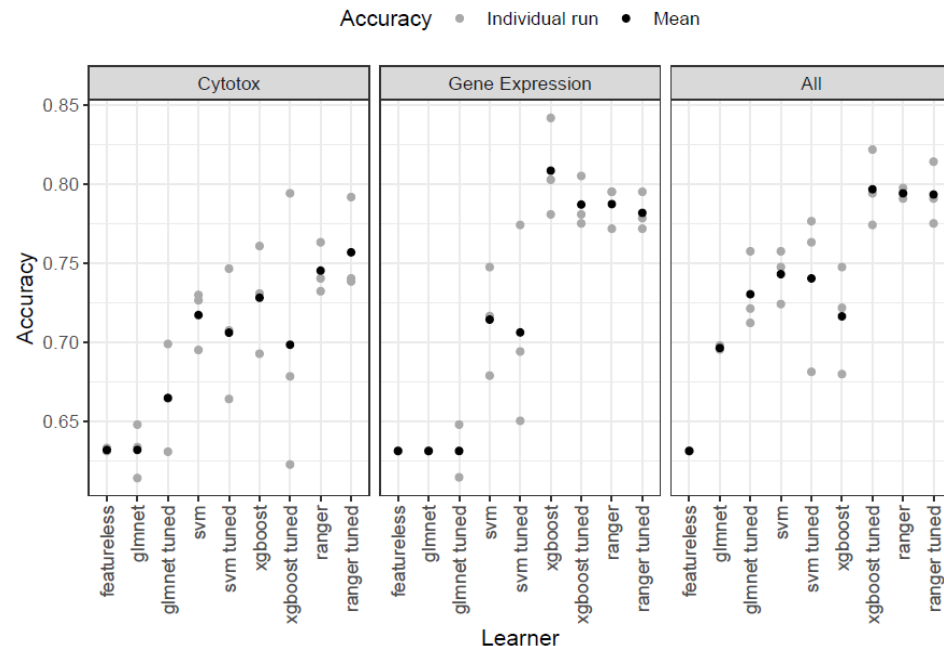


# Benefit of high-dimensional data

- Classification of hepatotoxicity of compounds based on cytotoxicity assays is improved by additional interpretable summaries of high-dimensional gene expression data

Marieke Stolte, Wiebke Albrecht, Tim Brecklinghaus, Lisa Gründler, Peng Chen, Jan G. Hengstler, Franziska Kappenberg, Jörg Rahnenführer.  
*Computational Toxicology* 28, 100288, 2023

- Goal: Prediction of hepatotoxicity
- Interpretable summaries of the gene expression data improve accuracy
- Summaries based on differentially expressed genes and fitting of parametric models

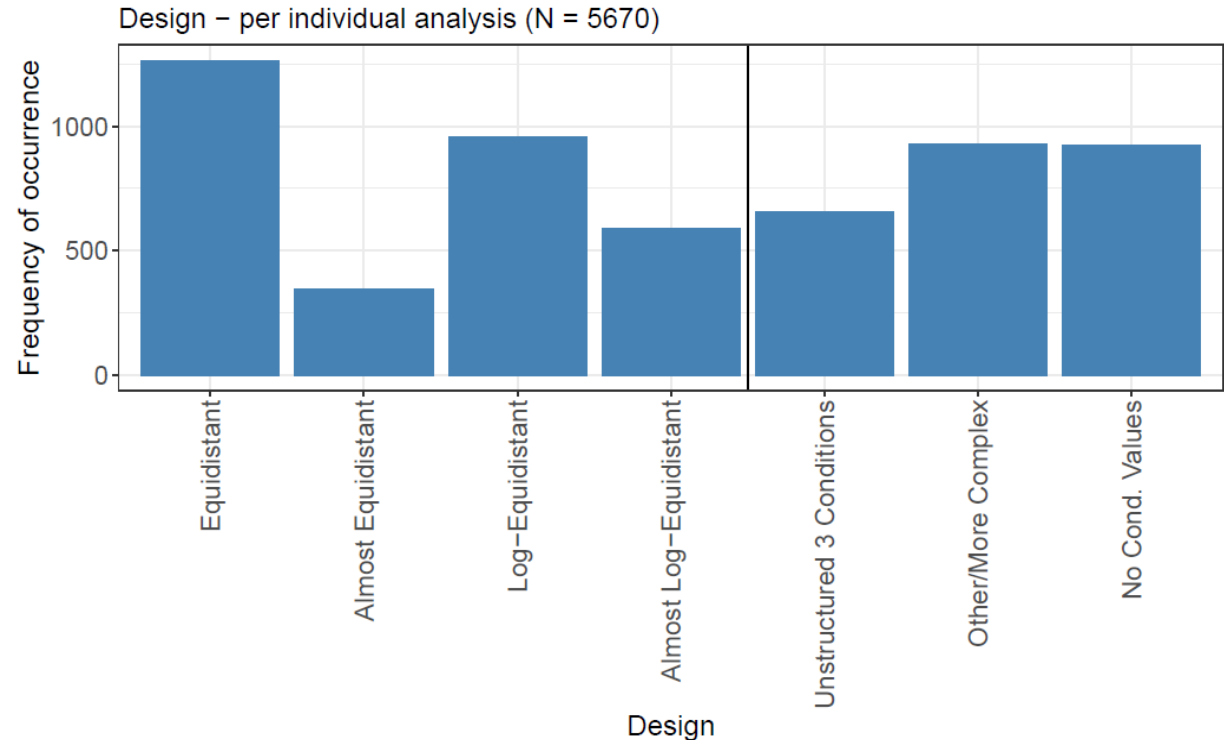


# Benefit of high-dimensional data

- Guidance for statistical design and analysis of toxicological dose-response experiments, based on a comprehensive literature review

Franziska Kappenberg et al., 2023

- **Statistical design**
- Results from the literature review with respect to the experimental design
- Mostly only equidistant or log-equidistant
  - Experimental reasons and lack of curve modelling



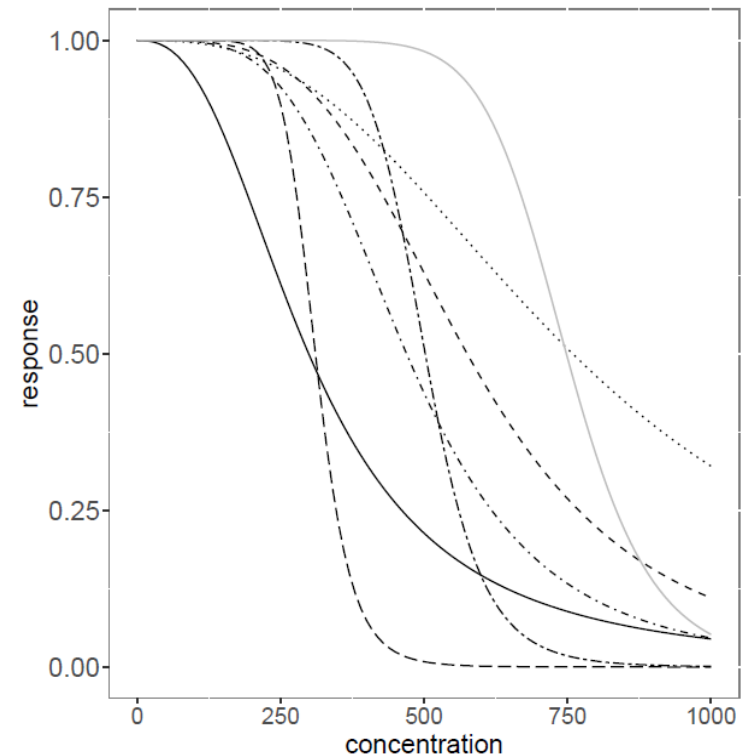
# Benefit of high-dimensional data

- **Design for the simultaneous inference of concentration-response curves**

Leonie Schürmeyer, Kirsten Schorning, Jörg Rahnenführer.

*BMC Bioinformatics*, 24, 393, 2023.

- **Statistical design for tens of thousands of genes**
- Efficient planning of design challenging, since same design (concentrations) for all genes
- Assume all concentration-response curves can be modelled with a suitable 4pLL model
- **Simultaneous D-optimal design:** Assume prior distribution over parameter space, determine design points using optimization algorithm



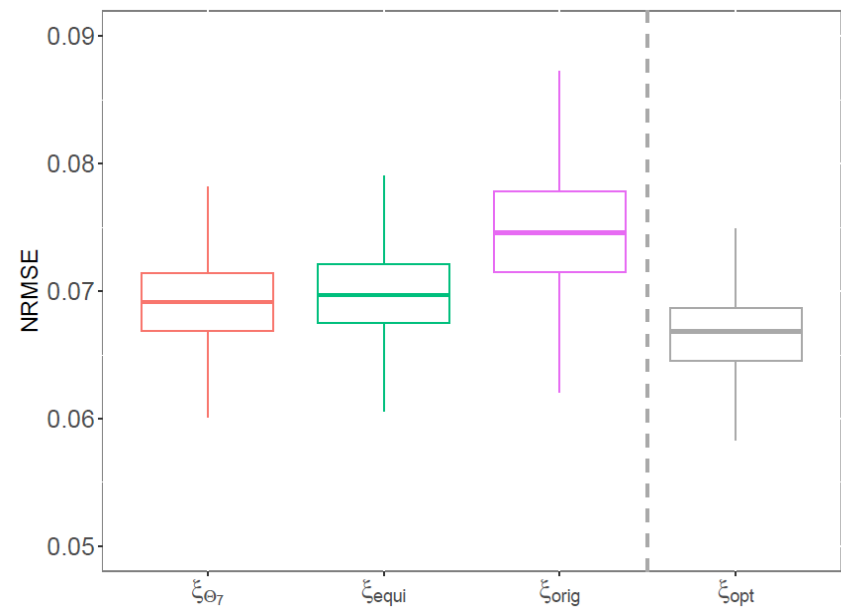
# Benefit of high-dimensional data

- **Design for the simultaneous inference of concentration-response curves**

Leonie Schürmeyer, Kirsten Schorning, Jörg Rahnenführer.

*BMC Bioinformatics*, 24, 393, 2023.

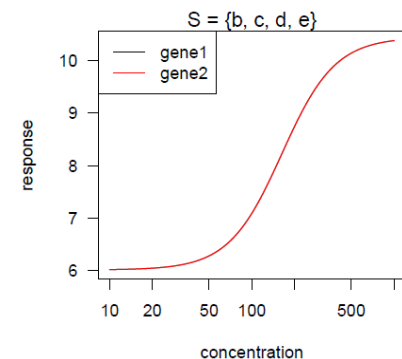
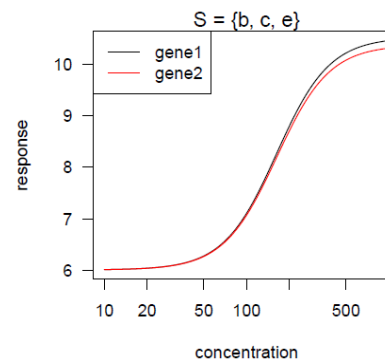
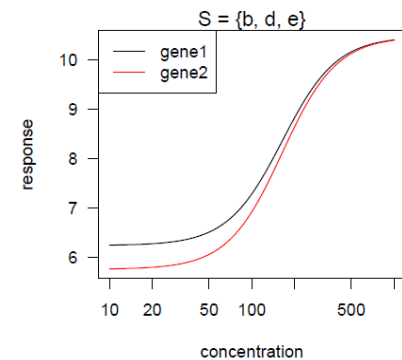
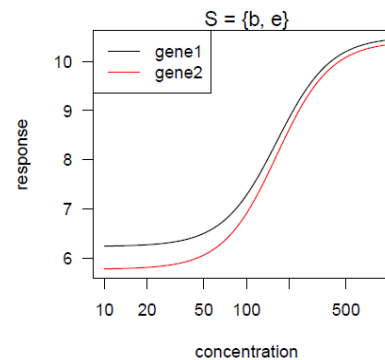
- **Statistical design for tens of thousands of genes**
- Results for comparison of designs in simulation study
  - RMSE compares estimated and true values of response in interval of interest
  - Bayesian design (red) is a little better than equidistant and clearly better than the originally used design



# Benefit of high-dimensional data

- **Parameter sharing for multiple dose-response curves**  
Onur Gül, Kirsten Schorning. Work in progress.

- Reduce number of parameters by fitting joint models for several genes (**common parameters**): bias/variance tradeoff !
- Focused information criterion for model selection to minimize the MSE for a specific parameter of the fitted curves
- Simulation results show a stronger decrease in MSE when using the focused information criterion in comparison to the AIC

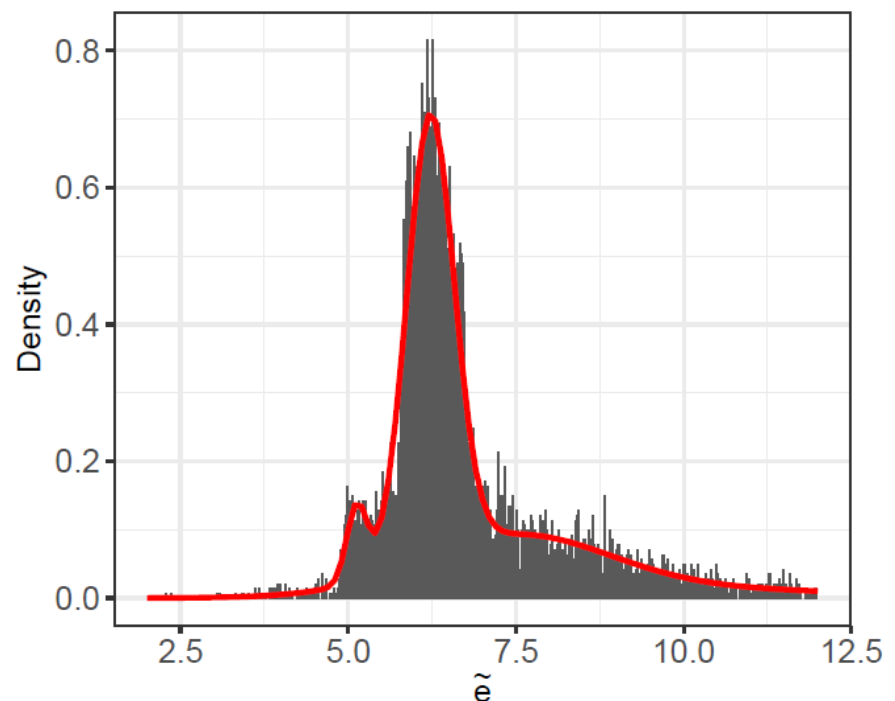


# Benefit of high-dimensional data

- Information sharing in high-dimensional gene expression data for improved parameter estimation in concentration-response modelling

Franziska Kappenberg, Jörg Rahnenführer. *PLOS One*, 18(10), e0293180. 2023

- Empirical Bayes approach for improving estimation of ED50 parameter of a parametric model
- Assume (mixture of) normal distributions as prior, then the posterior distribution is a (mixture of) normal distributions

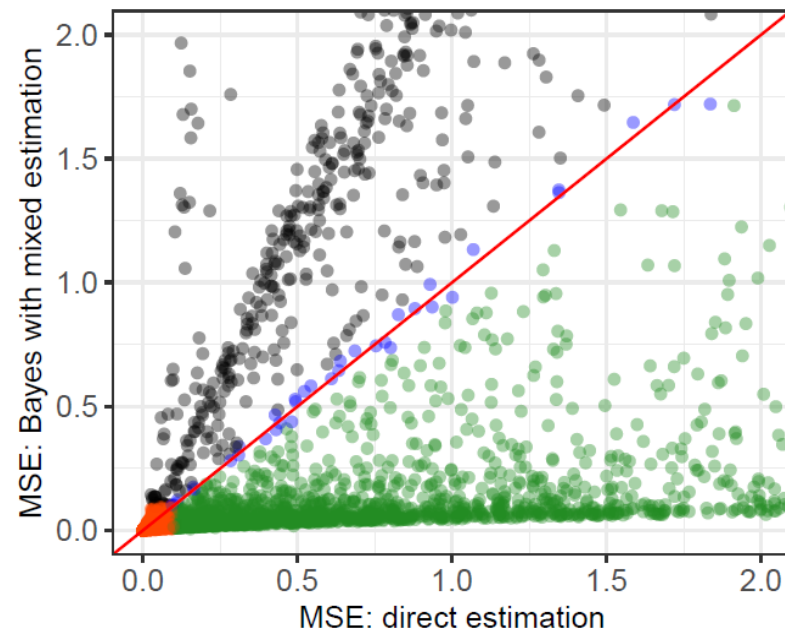


# Benefit of high-dimensional data

- Information sharing in high-dimensional gene expression data for improved parameter estimation in concentration-response modelling

Franziska Kappenberg, Jörg Rahnenführer. *PLoS One*, 18(10), e0293180. 2023

- Empirical Bayes approach for improving estimation of ED50 parameter of a parametric model
- Assume (mixture of) normal distributions as prior, then the posterior distribution is a (mixture of) normal distributions
- Plasmode simulation study: MSE clearly lower for many genes, but also higher for a few genes



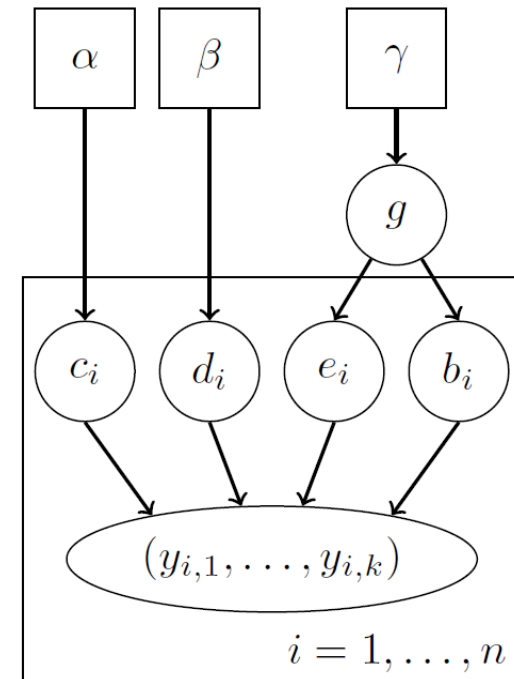


# Benefit of high-dimensional data

- Information sharing in high-dimensional gene expression data for improved parameter estimation in concentration-response modelling

Franziska Kappenberg, Jörg Rahnenführer. *PLoS One*, 18(10), e0293180. 2023

- Extension of the approach to fully Bayesian hierarchical model, inducing two-dimensional shrinkage
- Weakly informative functionally uniform priors for the 4pLL parameters
- Model fitted to subset of genes with approximately monotone profiles



# Acknowledgements

## RTG 2624 – Open PhD positions!



Dr.  
Franziska  
Kappenberg



Dr. Tamara  
Schikowski,  
Deputy  
spokesperson



M.Sc.  
Julia  
Duda



M.Sc.  
Leonie  
Schürmeyer

# Benefit of high-dimensional data

- Model selection characteristics when using MCP-Mod for dose–response gene expression data

Julia Duda, Franziska Kappenberg, Jörg Rahnenführer. *Biometrical Journal*, 64(5), 883-897, 2022

- Multiple Comparison Procedure (MCP) and Modelling, two-step procedure originally developed for Phase II clinical studies
- ‘Guesstimates’ for candidate models are required for all genes simultaneously
- With higher signal-to-noise ratio, the sigE<sub>max</sub> model wins more often

