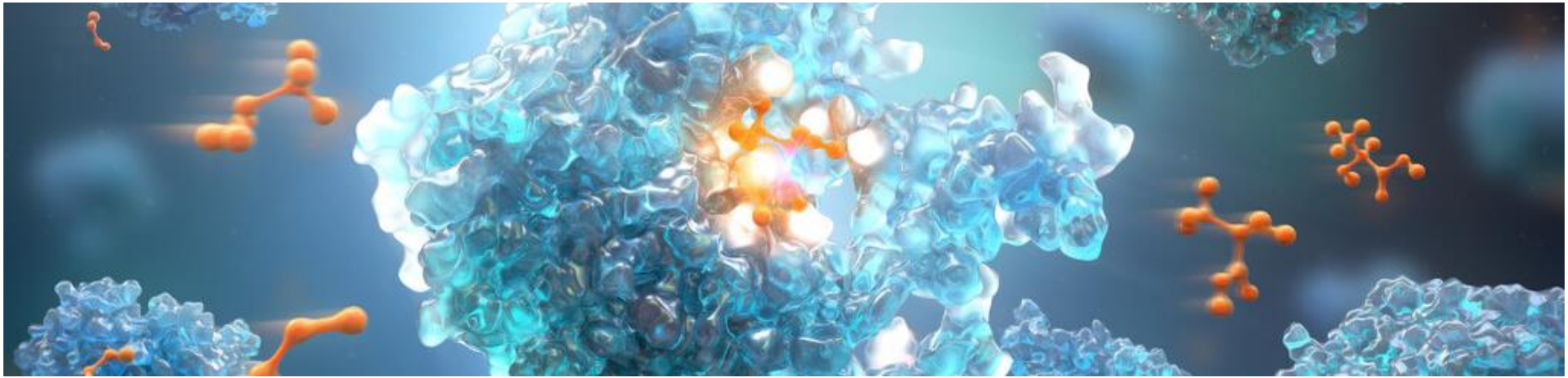
Bayesian neural networks and **changepoint Gaussian Processes** for drug safety and efficacy: two case studies

Elizaveta Semenova

Imperial College London, School of Public Health

Stan for Pharmacometrics, INSERM

4 July 2025



Context

- Imperial College London, Assistant Professor, 2024 – now
- University of Oxford, Schmidt Sciences AI2050 Early Career Fellow, 2023 – 2024
- University of Oxford, Postdoc, 2022 – 2023
- Imperial College London, Postdoc, 2021 – 2022
- AstraZeneca, Postdoc, 2019 – 2021
- Swiss Tropical and Public Health Institute, PhD, 2014 – 2019

Context

- Interests:
  - applied Bayesian inference
  - spatial statistics
  - epidemiology (both infectious and NCDs)
  - deep generative models

Bayesian Neural Networks for toxicity prediction

Computational Toxicology 16 (2020) 100133

Contents lists available at ScienceDirect

Computational Toxicology

journal homepage: www.elsevier.com/locate/comtox

A Bayesian neural network for toxicity prediction

Elizaveta Semenova[a,*], Dominic P. Williams[b], Avid M. Afzal[a], Stanley E. Lazic[c]

[a] Data Sciences and Quantitative Biology, Discovery Sciences, R&D, AstraZeneca, Cambridge, UK
[b] Functional and Mechanistic Safety, Clinical Pharmacology and Safety Sciences, R&D, AstraZeneca, Cambridge, UK
[c] Prioris.ai Inc, Ottawa, Canada

DILI

- Drug-induced liver injury (DILI) is a major cause of attrition in drug development and a common reason for withdrawing a drug from the market.

DILI

- Drug-induced liver injury (DILI) is a major cause of attrition in drug development and a common reason for withdrawing a drug from the market.

- Predicting clinical DILI is difficult due to its multi-mechanistic nature and chemical properties of the drug.

DILI

- Drug-induced liver injury (DILI) is a major cause of attrition in drug development and a common reason for withdrawing a drug from the market.
- Predicting clinical DILI is difficult due to its multi-mechanistic nature and chemical properties of the drug.
- Pre-clinical animal studies fail in making correct predictions in about 45% of clinical trials*.

* Concordance of the toxicity of pharmaceuticals in humans and in animals, Olson et al., Regulatory Toxicology and Pharmacology (2000)

DILI

- Drug-induced liver injury (DILI) is a major cause of attrition in drug development and a common reason for withdrawing a drug from the market.
- Predicting clinical DILI is difficult due to its multi-mechanistic nature and chemical properties of the drug.
- Pre-clinical animal studies fail in making correct predictions in about 45% of clinical trials*.
- Classical *in silico* models require sufficient amount of data to make reliable predictions, while real life liver toxicity data sets are small.

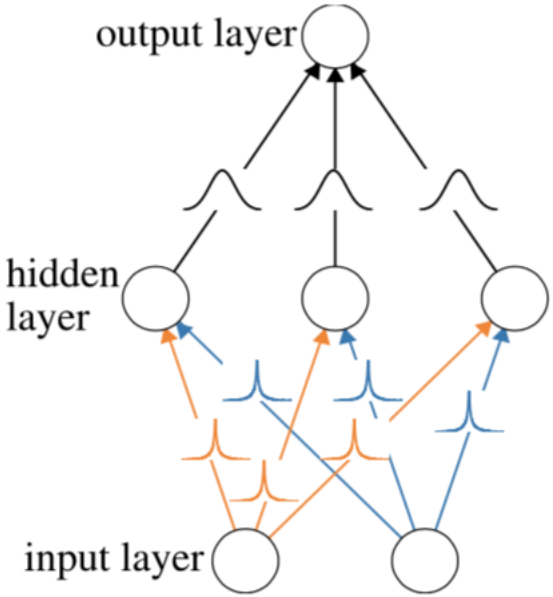
* Concordance of the toxicity of pharmaceuticals in humans and in animals, Olson et al., Regulatory Toxicology and Pharmacology (2000)

Neural Networks for toxicity prediction

- Neural networks (NNs) are popular due to their flexibility.
- NNs have been applied in the context of DILI prediction*.
- However, they are not recommended for small data sets and do not provide a degree of uncertainty on their predictions.

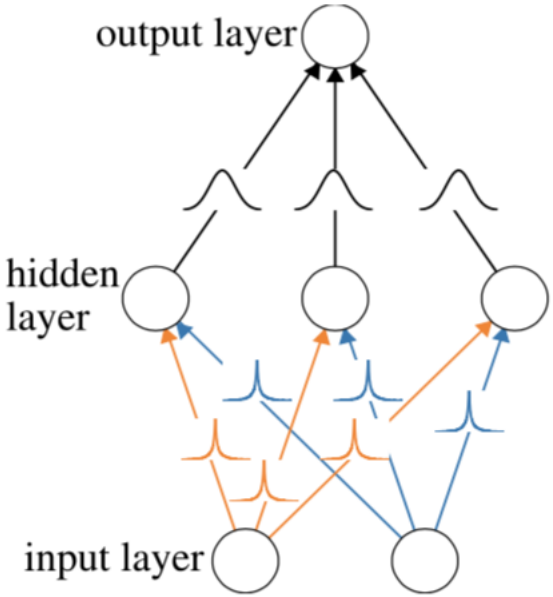
*Deep learning for drug-induced liver injury, Xu et al. Journal of Chemical Information and Modelling (2015)

Bayesian Neural Networks



- Bayesian neural networks (BNNs) describe parameters of a NN model via distributions, rather than a single number.

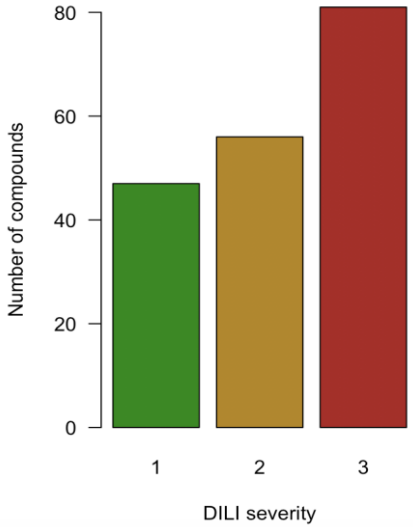
Bayesian Neural Networks



- Bayesian neural networks (BNNs) describe parameters of a NN model via distributions, rather than a single number.

- Bayesian models
  - prevent overfitting by using prior distributions
  - provide information about the degree of uncertainty of predictions

Data



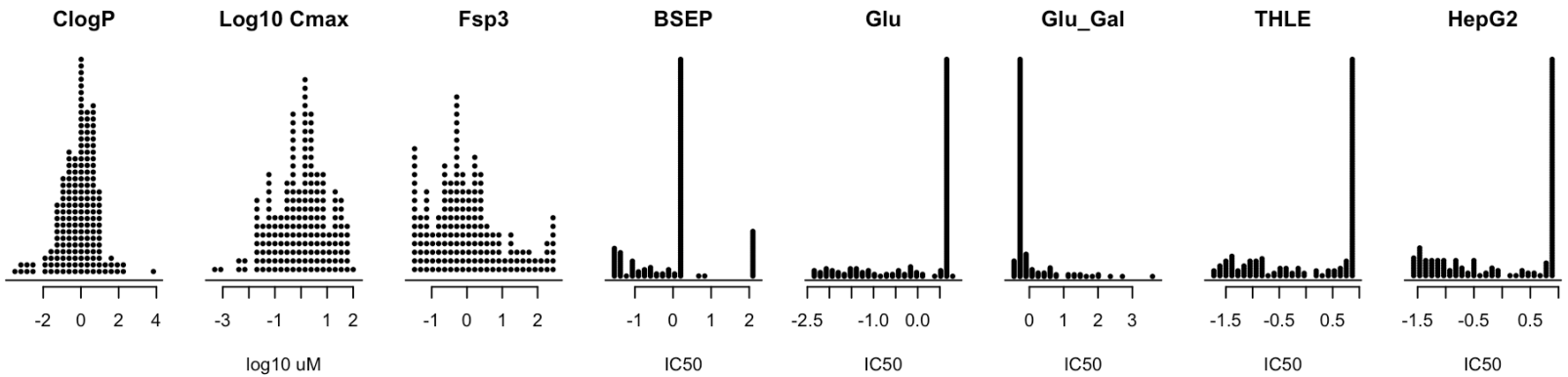
1 – no DILI concern
2 – less DILI concern
3 – most DILI concern

- We used the dataset provided in Aleo et al* containing 184 labelled compounds.
- A random train-test 80% - 20% split was created exactly matching proportions of each severity category.

*Moving beyond Binary Predictions of Human Drug-Induced Liver Injury (DILI) toward Contrasting Relative Risk Potential, Aleo et al. Chemical Research in Toxicology (2019)

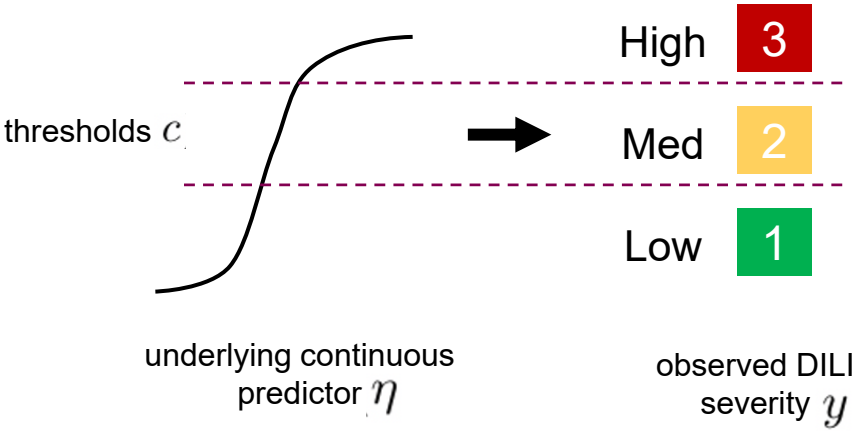
Data

- Predictors: assays and physicochemical properties of compounds



Models

- DILI severity was modelled via the ordered logistic regression with three classes.
- Thresholds, separating the classes, were estimated from data:

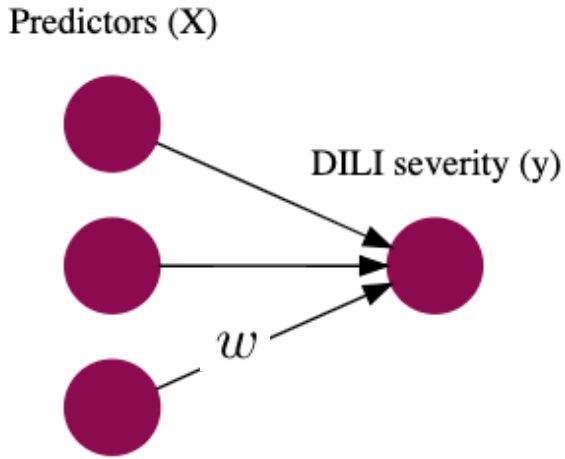


$$y \sim \text{OrderedLogistic}(\eta, c)$$

Models

Proportional odds logistic regression (POLR)

- Model structure*



- Priors

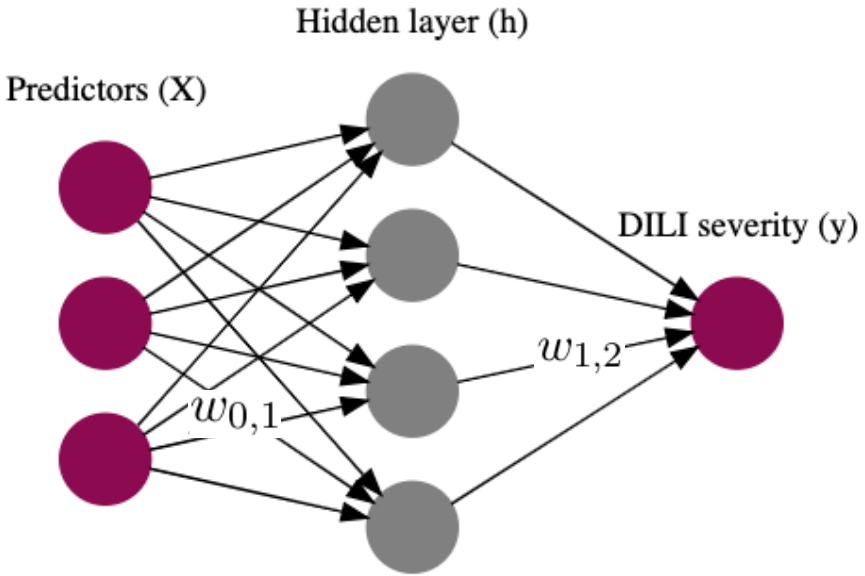
$$\eta = Xw$$
$$w \sim \text{Normal}(0, \sigma)$$
$$\sigma \sim \text{Normal}^{+}(0, 1)$$

* Predicting drug-induced liver injury with Bayesian Machine Learning. Williams D., Lazic S. et al. Chemical Research in Toxicology (2019)

Models

Bayesian Neural Network (BNN)

- Model structure



- Priors

$$h = \mathrm{ReLU}(X w_{0,1})$$
$$\eta = h w_{1,2}$$
$$w = [w_{0,1}, w_{1,2}]$$
$$w \sim \mathrm{Normal}(0, \sigma)$$
$$\sigma \sim \mathrm{Normal}^{+}(0, 1)$$

Evaluation metrics

- **WAIC** – *Watanabe-Akaike Information criterion* for Bayesian model selection, applicable to models with non-normal posteriors (the smaller, the better)

Evaluation metrics

- **WAIC** – *Watanabe-Akaike Information criterion* for Bayesian model selection, applicable to models with non-normal posteriors (the smaller, the better)

- **OBS** – *Ordered Brier Score* measures the distance from predicted probability to the true class, accounting for the ordered nature of the data; this measure is more suitable than balanced accuracy for ordered outcomes (the smaller, the better)

Evaluation metrics

- **WAIC** – *Watanabe-Akaike Information criterion* for Bayesian model selection, applicable to models with non-normal posteriors (the smaller, the better)

- **OBS** – *Ordered Brier Score* measures the distance from predicted probability to the true class, accounting for the ordered nature of the data; this measure is more suitable than balanced accuracy for ordered outcomes (the smaller, the better)

- **BSS** – *Brier Skill Score* measures how much better a model is than the baseline model predicting observed frequencies (the larger, the better)

Evaluation metrics

- **WAIC** – *Watanabe-Akaike Information criterion* for Bayesian model selection, applicable to models with non-normal posteriors (the smaller, the better)

- **OBS** – *Ordered Brier Score* measures the distance from predicted probability to the true class, accounting for the ordered nature of the data; this measure is more suitable than balanced accuracy for ordered outcomes (the smaller, the better)

- **BSS** – *Brier Skill Score* measures how much better a model is than the baseline model predicting observed frequencies (the larger, the better)

- **BA** – *Balanced accuracy* takes imbalances in the observed outcome into account (the larger, the better)

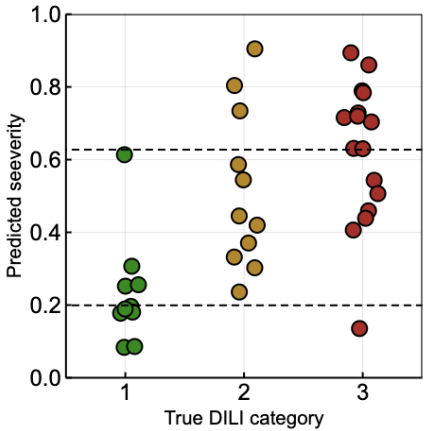
Results

Comparison of models according to several evaluation metrics

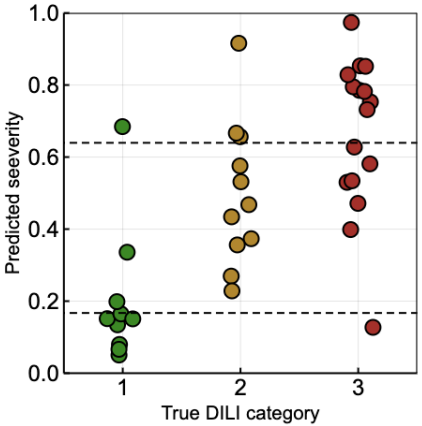
Model	WAIC	Train / Test mean OBS	Train / Test median OBS	Train / Test mean BSS	Train / Test median BSS	Train / Test BA
POLR	272.0	0.14 / 0.16	0.10 / 0.12	0.27 / 0.19	0.38 / 0.35	0.64 / 0.62
BNN	253.1	0.12 / 0.14	0.08 / 0.10	0.36 / 0.29	0.45 / 0.37	0.71 / 0.67

Predictions for test examples

- BNN displays sharper separation between categories



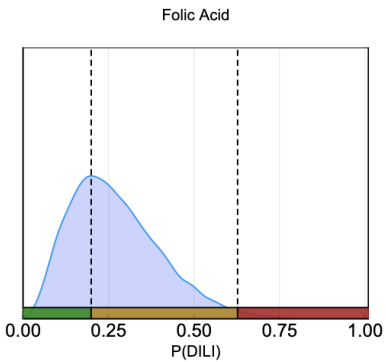
POLR

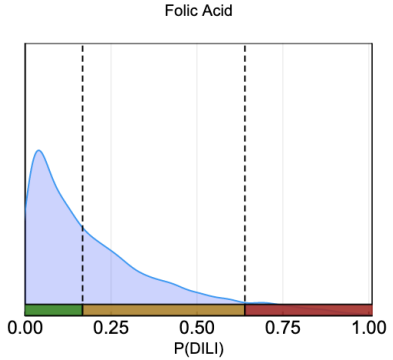


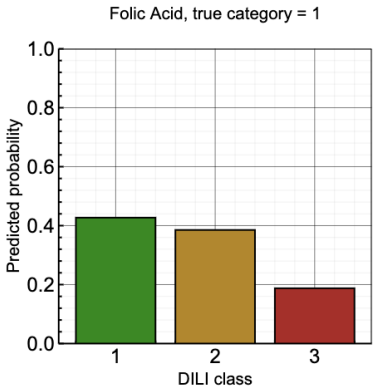
BNN

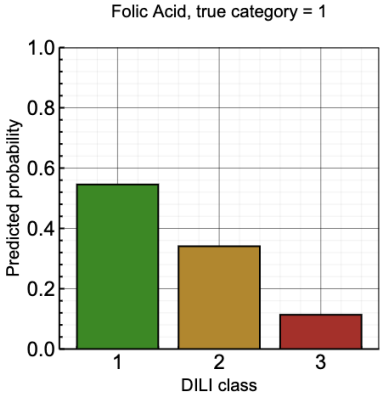
Predictions for one compound

- Posterior distributions
- Predictions









POLR

BNN

Conclusions

- In our application the BNN performs better than a traditional but less flexible POLR model with interactions and does not show strong signs of overfitting on a relatively small dataset.

- We provide the first application of BNNs to toxicology.

- The presented model lays a foundation for more complex models built on larger datasets but can already be adopted by safety pharmacologists for risk quantification.

Changepoint Gaussian Processes for drug efficacy

Original Research

Flexible Fitting of PROTAC Concentration–Response Curves with Changepoint Gaussian Processes

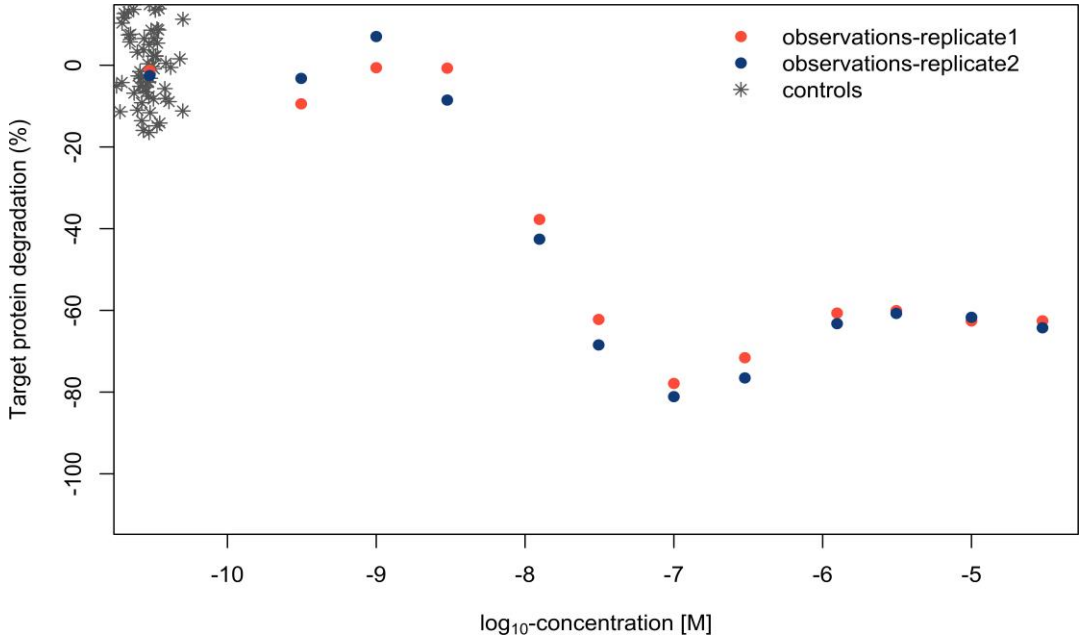
SLAS Discovery
2021, Vol. 26(9) 1212–1224
© Society for Laboratory Automation and Screening 2021
DOI: 10.1177/24725552211028142
journals.sagepub.com/home/jbx
SAGE

Elizaveta Semenova[1], Maria Luisa Guerriero[1], Bairu Zhang[1], Andreas Hock[2], Philip Hopcroft[2], Ganesh Kadamur[2], Avid M. Afzal[1], and Stanley E. Lazic[3]

Introduction

- **Concentration-response** (CR) experiments are used to rank drug candidates.
- Traditional small molecules typically yield **sigmoidal curves**, characterized by a **plateau** at high drug concentrations.
- CR curves of a **new drug modality** show a loss of efficacy at higher doses, known as the **‘hook effect’.**

Understanding data



Domain understanding

We are looking to fit a curve which is

- flat at **low concentrations** (no compound activity),
- able to capture curve characteristics at **higher concentrations** (the ‘hook effect’).

Sources of uncertainty

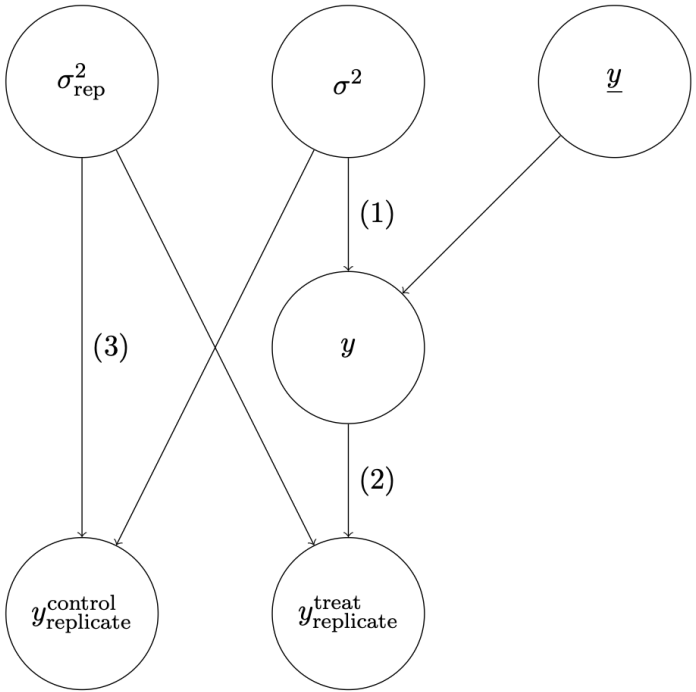
We account for two sources of uncertainty:

- **curve uncertainty**,
- **replicate-to-replicate** variation**.**

$$y \sim N(\underline{y}, \sigma^2 I), \tag{1}$$

$$y_{\text{rep}}^{\text{treatment}} \sim t_\nu(y, \sigma_{\text{rep}}), \tag{2}$$

$$y_{\text{rep}}^{\text{control}} \sim t_\nu(\mu, \sqrt{\sigma^2 + \sigma_{\text{rep}}}) \tag{3}$$

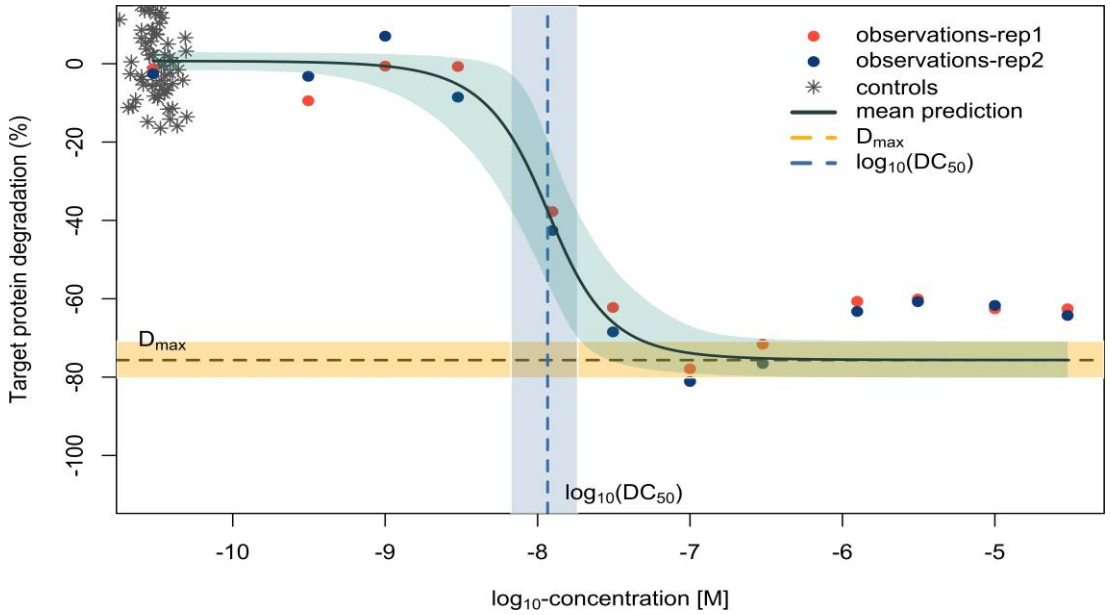


Traditional Hill's (4PL) model

$$\underline{y}(x) = d + \frac{a - d}{1 + \exp^{(-b(x-c))}}$$

- d: degradation at **zero concentration**,
- a: D_{max} - **maximal degradation**,
- c: $\log_{10}(DC_{50})$ - concentration of **half-degradation**,
- x: dose on the $\log_{10}$-scale,
- b: **Hill's slope** (slope at the half-degradation point).

Model fit


observations-rep1
observations-rep2
controls
mean prediction
D_{max}
$\log_{10}(DC_{50})$
Target protein degradation (%)
0
-20
-40
-60
-80
-100
D_{max}
$\log_{10}(DC_{50})$
-10
-9
-8
-7
-6
-5
$\log_{10}$-concentration [M]

Gaussian Process (GP) model

Gaussian Process model allows to fit flexible curve shapes. It is defined as

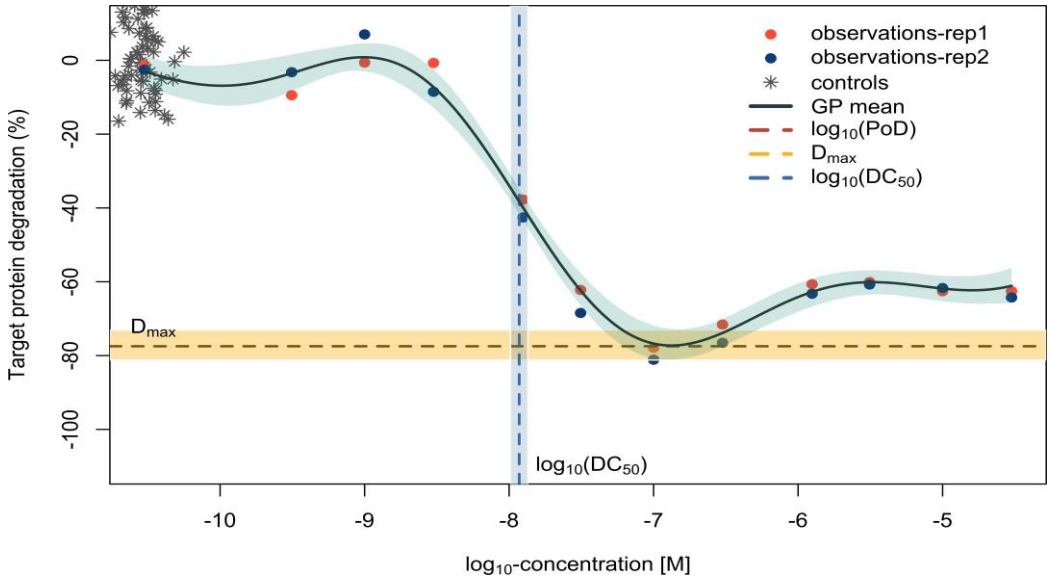
$$f \sim \mathrm{GP}(0, k).$$

Evaluated on a finite set of points it constitutes a multivariate normal with covariance matrix K. For example

$$K[i,j] = \eta^2 \exp\left(-\frac{(x_i - x_j)^2}{\rho^2}\right).$$

Parameters η and ρ define the **amplitude** and **lengthscale** of the curve, correspondingly.

Model fit


observations-rep1
observations-rep2
controls
GP mean
log10(PoD)
Dmax
log10(DC50)
Target protein degradation (%)
0
-20
-40
-60
-80
-100
Dmax
log10(DC50)
-10
-9
-8
-7
-6
-5
log10-concentration [M]

Changepoint Gaussian Process

Kernel design allows to specify a wider range of GP priors. Given two GPs

$$f_1(x) \sim \mathrm{GP}(0, k_1),$$
$$f_2(x) \sim \mathrm{GP}(0, k_2),$$

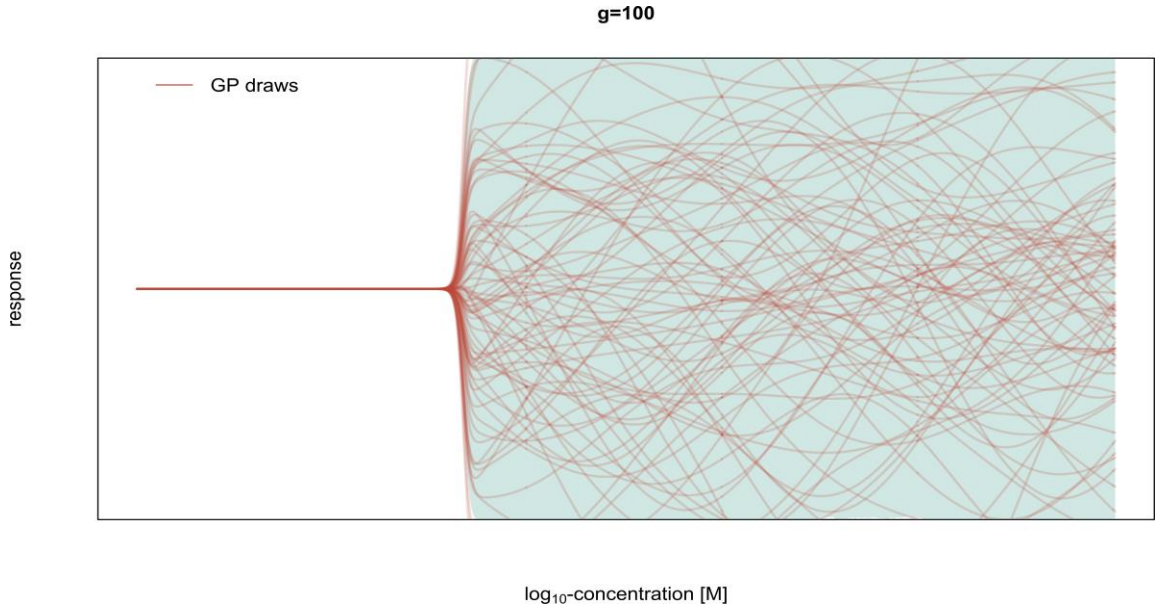
we can construct a new one

$$f_\theta(x) = (1 - w_\theta(x))f_1(x) + w_\theta(x)f_2(x),$$
$$w_\theta(x) = \sigma(g(x - \theta)), g > 1.$$

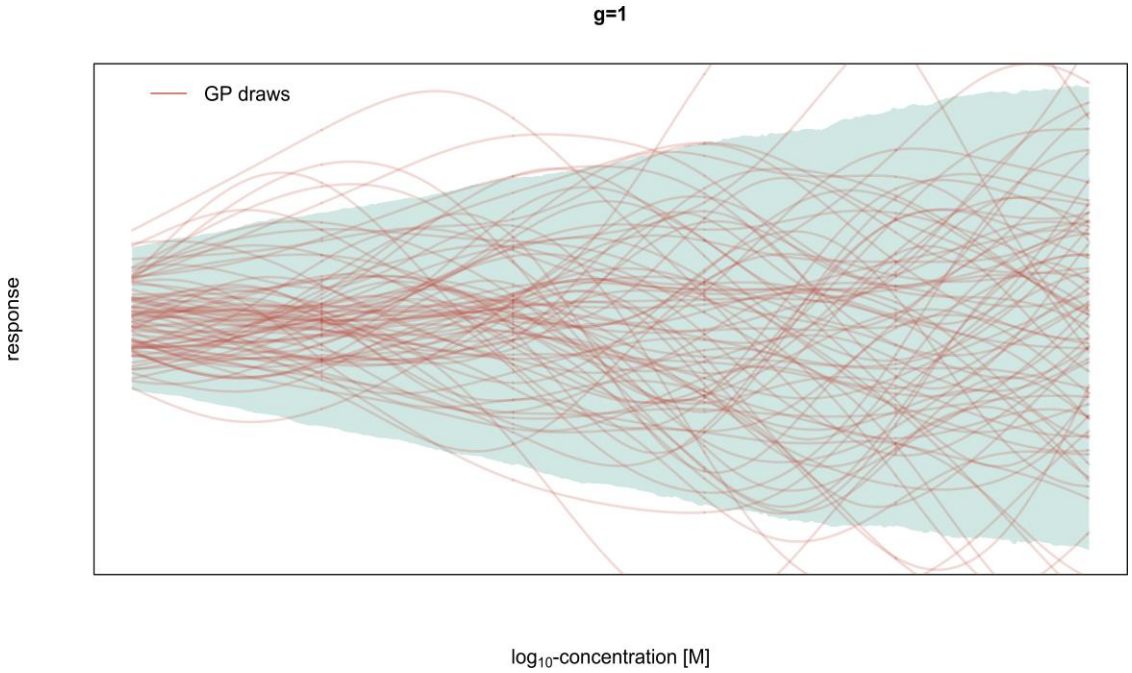
Then

$$f_\theta(x) \sim \mathrm{GP}(0, k_\theta),$$
$$k_\theta(x, x') = (1 - w_\theta(x))k_1(x, x')(1 - w_\theta(x')) + w_\theta(x)k_2(x, x')w_\theta(x').$$

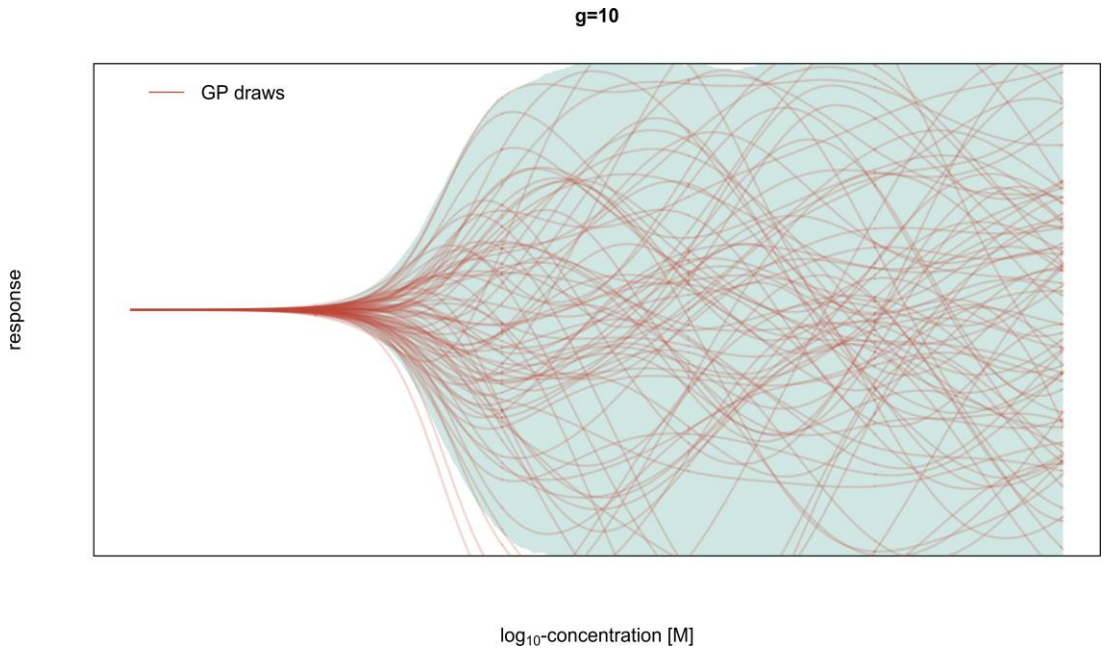
Changepoint GP priors



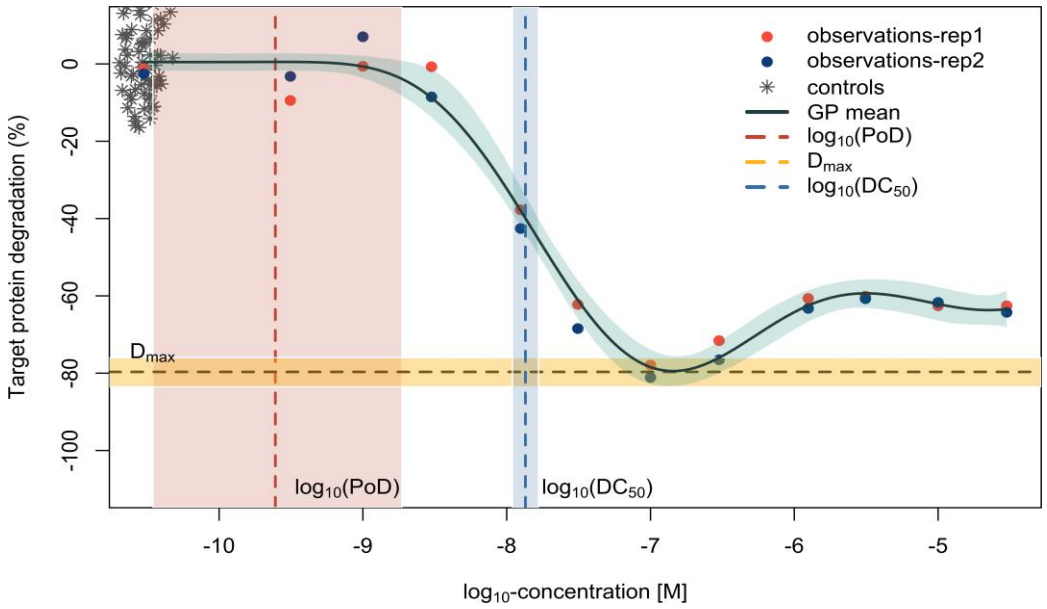
Changepoint GP priors



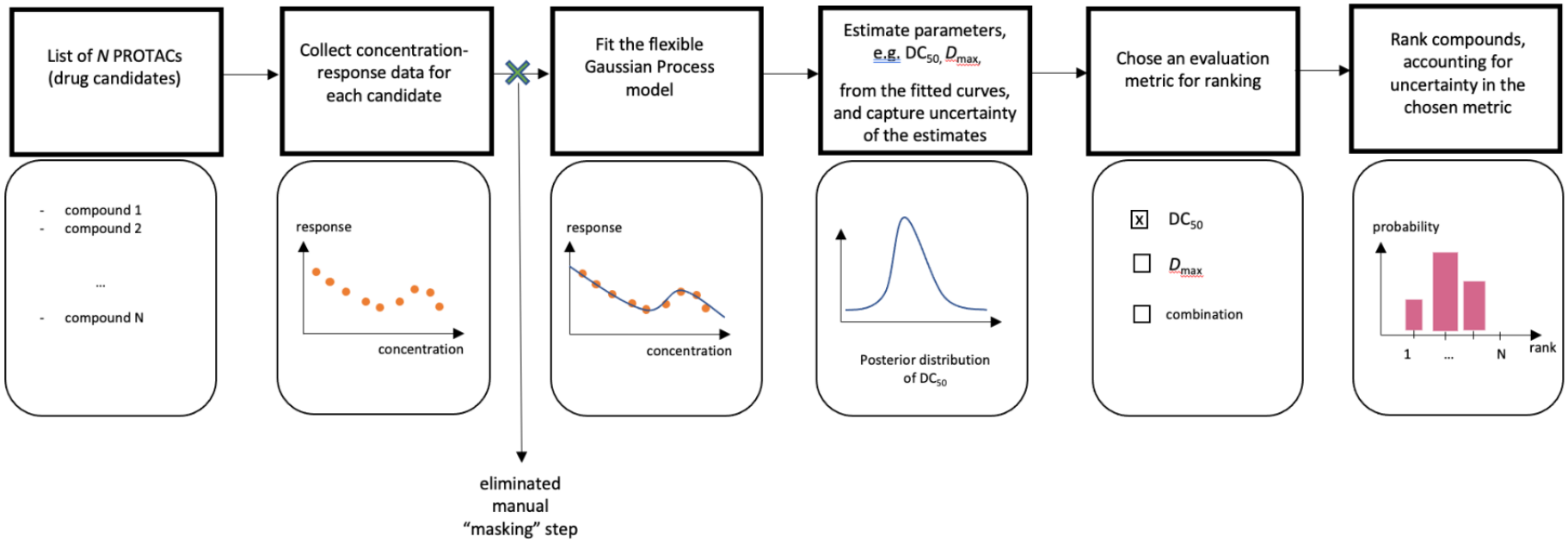
Changepoint GP priors



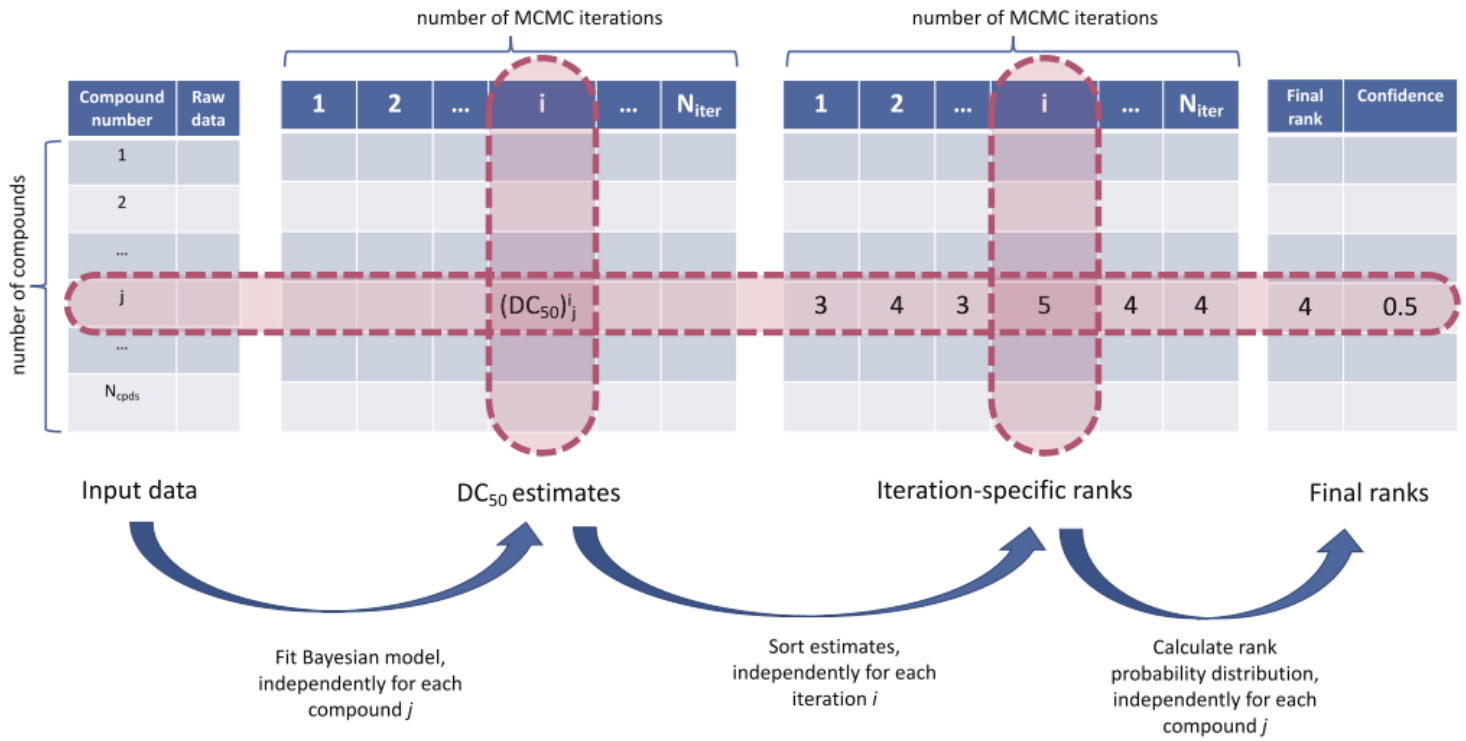
Model fit


observations-rep1
observations-rep2
controls
GP mean
log10(PoD)
Dmax
log10(DC50)
Target protein degradation (%)
0
-20
-40
-60
-80
-100
Dmax
log10(PoD)
log10(DC50)
-10
-9
-8
-7
-6
-5
log10-concentration [M]

The process of compound ranking



From raw data to compound ranks



Thank you

And stay in touch:

elizaveta.p.semenova@gmail.com