BDA Project on “QSAR aquatic toxicity”

HOANG DUNG PHAM NGUYEN (899376),

FRANCESCO ROTA (887032),

PRAGATI GUPTA (881533)

Introduction

- Waters on Earth more and more poisoned
- Toxicity measured as lethality over animals
- QSAR – Quantitative Structure Activity Relationship

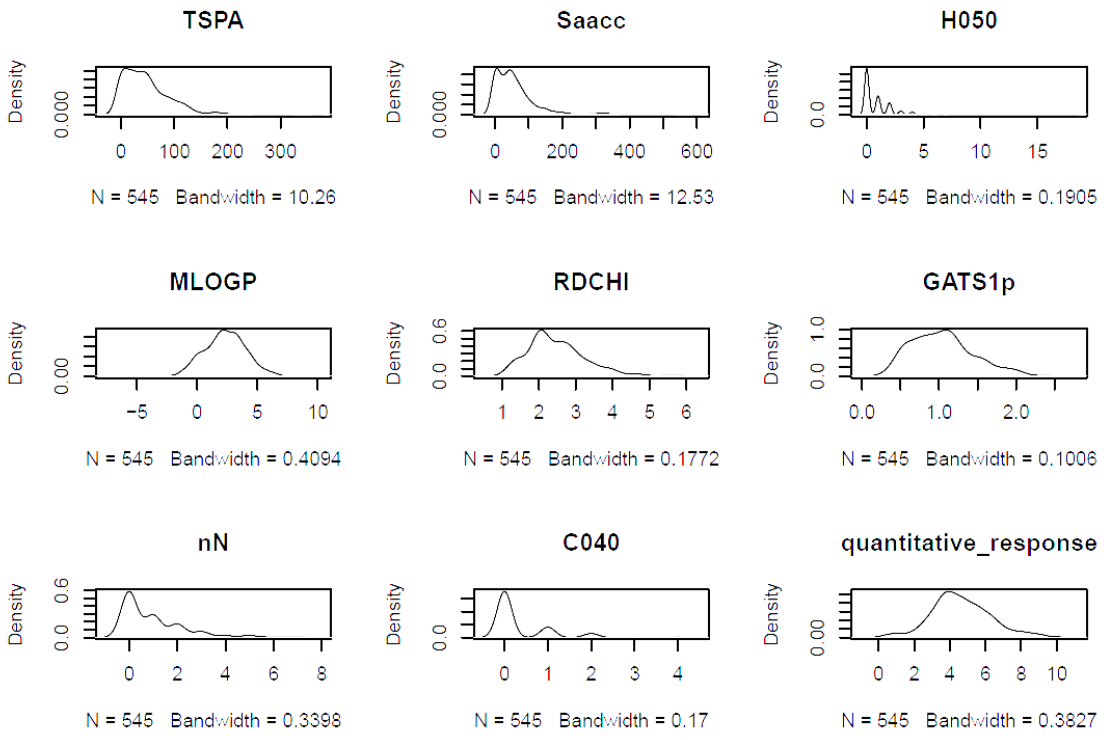
Daphia Magna

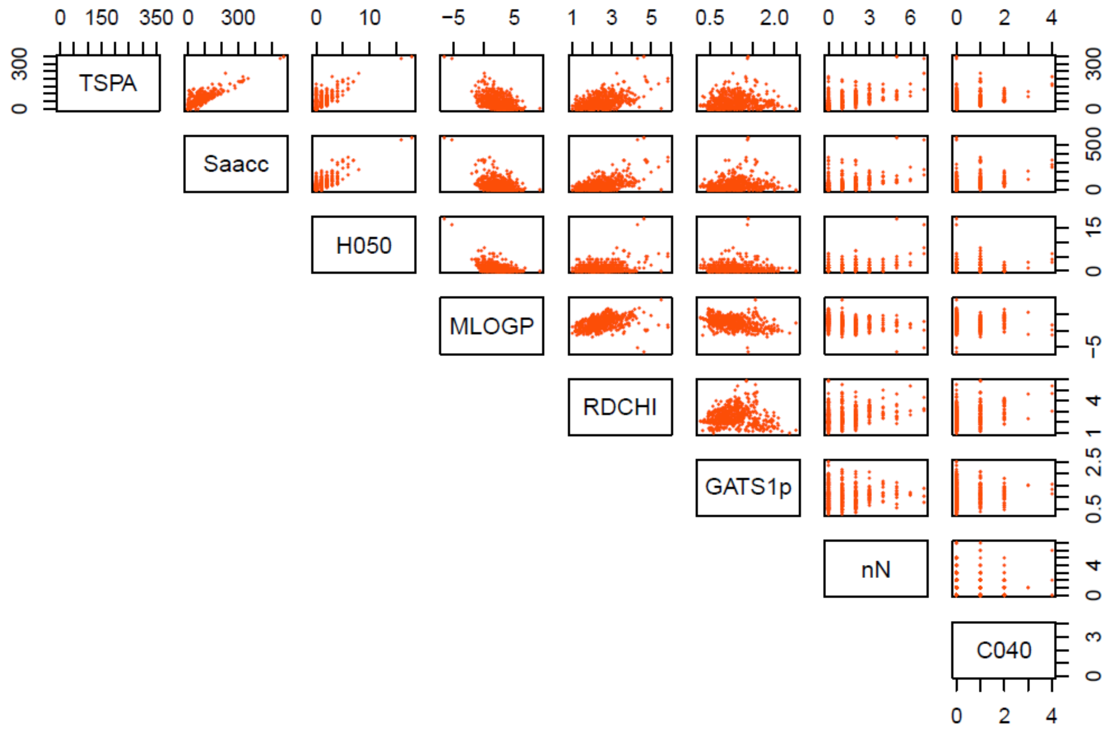
Dataset

- The predictive value (**Quantitative Response**) is the level of LC50 that we want to conclude from the dataset.
- Measurement LC50 is survivability of Daphia Magna
- Can predict the LC50 level given molecule information?
- 426 chemicals → 8 molecular descriptors

feature number	Feature name	Feature Description
1	TSPA	Tot Molecular properties
2	SAACC	Molecular properties
3	H050	Atom-centred fragments
4	MLOGP	Molecular properties
5	RDCHI	Connectivity indices
6	GATS1p	2D autocorrelations
7	nN	Constitutional indices
8	C040	Atom-centred fragments
9	Quantitative Response	acute aquatic toxicity

Analysis of the Dataset





Models

Linear Model

- Best choice as per previous study for Frequentist Approach

Hierarchical Model

- As per study, number of carbon and hydrogen atom may divide the dataset in subgroups

Gaussian Processes

- It gives you a possibility to go into infinite dimension
- Kernel process

Prior Selection

Linear Model

$$\sigma \sim \mathcal{N}(0, 10)$$

$$W \sim \mathcal{N}(0, 1)$$

Hierarchical Model

$$\sigma \sim \mathcal{N}(0, 10)$$

$$\mu_c \sim \mathcal{N}(0, 1)$$

$$\sigma_c \sim \mathcal{N}(0, 10)$$

Gaussian Process

$$\rho \sim InvGamma(5, 5)$$

$$\alpha \sim \mathcal{N}(0, 1)$$

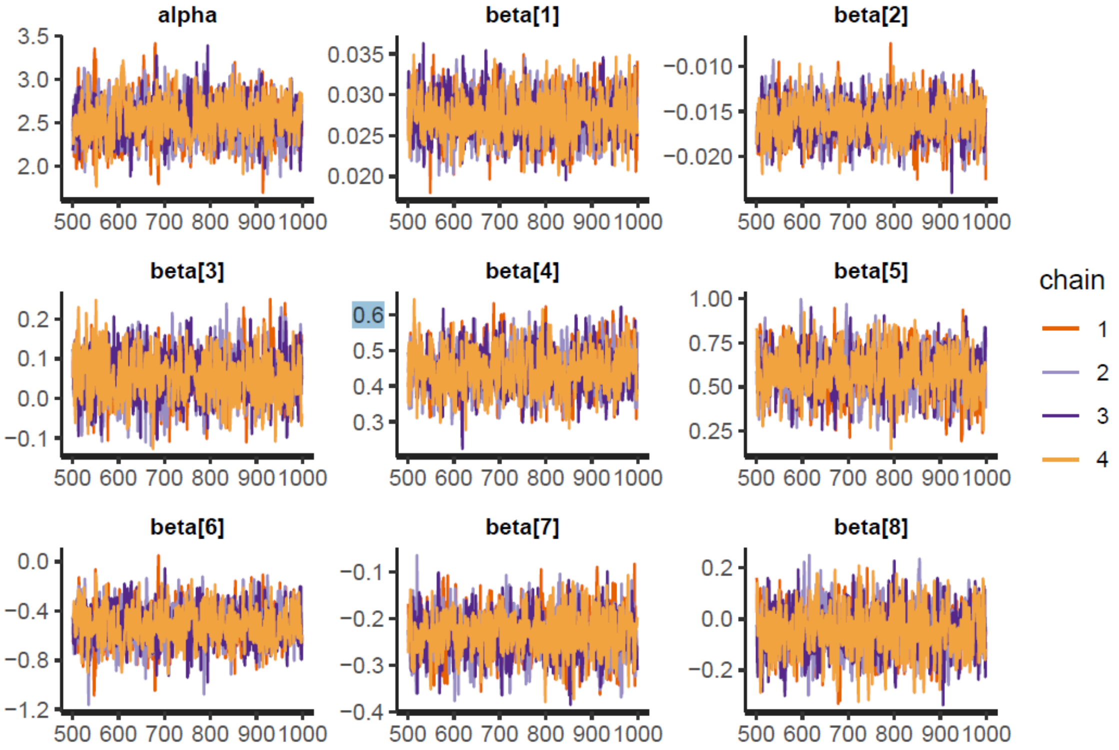
$$\sigma \sim \mathcal{N}(0, 1)$$

$$\eta \sim \mathcal{N}(0, 1)$$

https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations

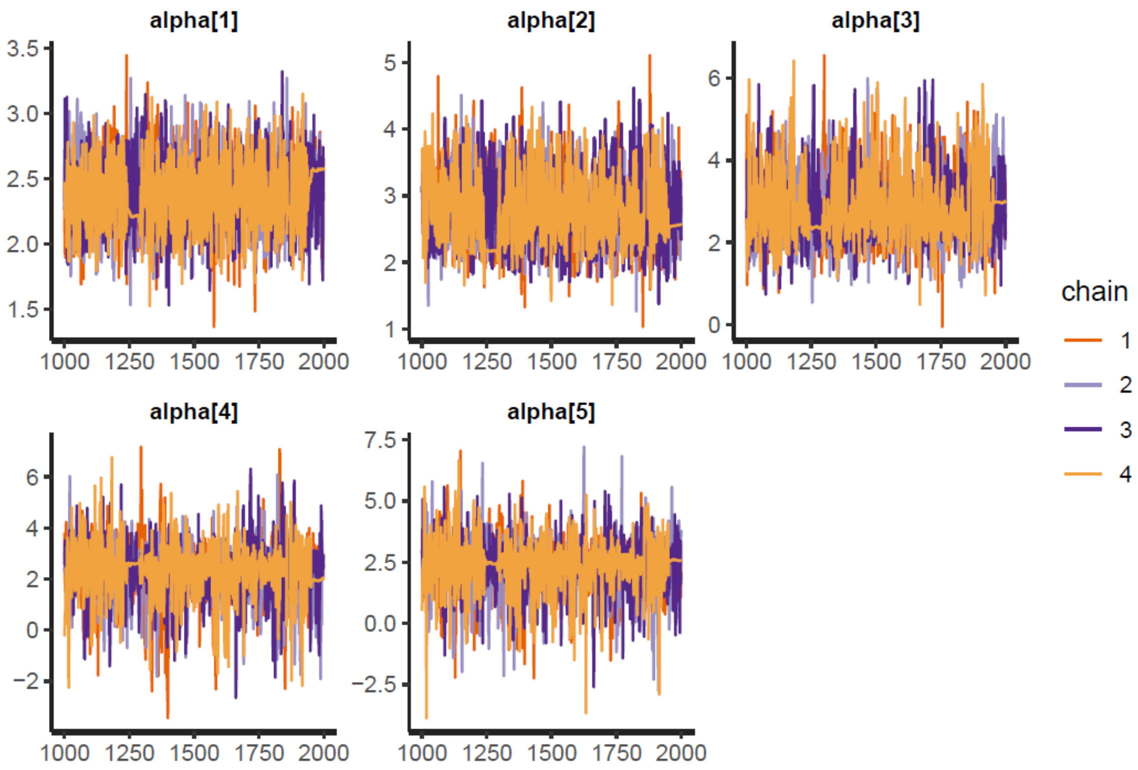
Linear Model

- Model $y \sim \mathcal{N}(\mu, \sigma)$
- where $\mu = W \times X$
- Stan model has been run for 1000 iterations, 4 Markov chains
- divided into train and test (20 samples) sets
- $\hat{R}$ = 1.004526
- n_{eff} > 0.01 for all parameters
- Bulk ESS over 100 for all the parameters
- No divergences in the linear model



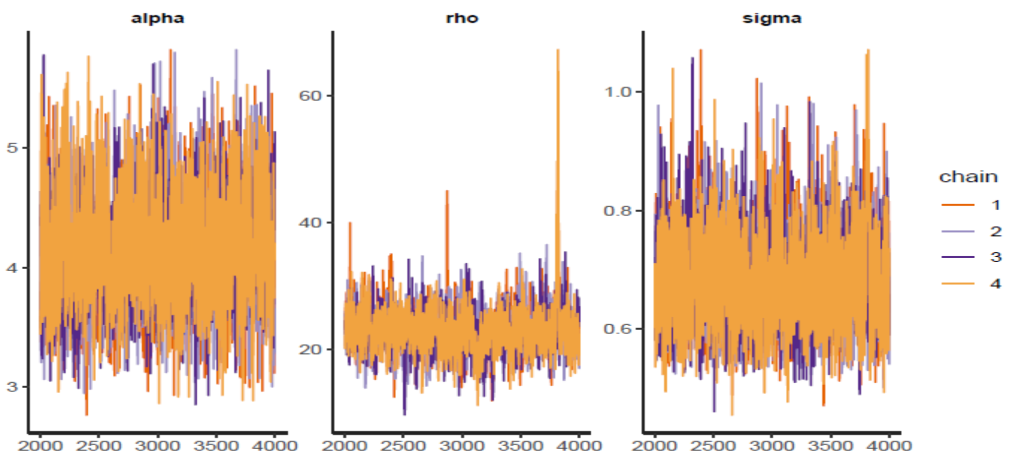
Hierarchical Model

- Model $y_{ij} \sim \mathcal{N}(\mu_i, \sigma)$
- where $\mu_i = W_i \times X \qquad W_i \sim \mathcal{N}(u_c, \sigma_c)$
- Stan model has been run for 2000 iterations, 4 Markov chains
- adapt_delta was set to 0.95
- $\hat{R}$ = 1.01686
- n_{eff} > 0.01 for all parameters
- Bulk ESS less than 100 for some parameters
- 239 divergences in the model

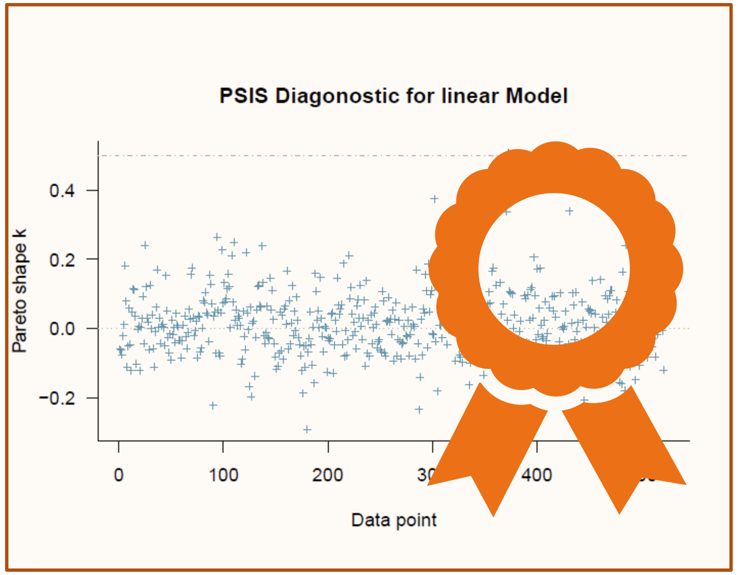


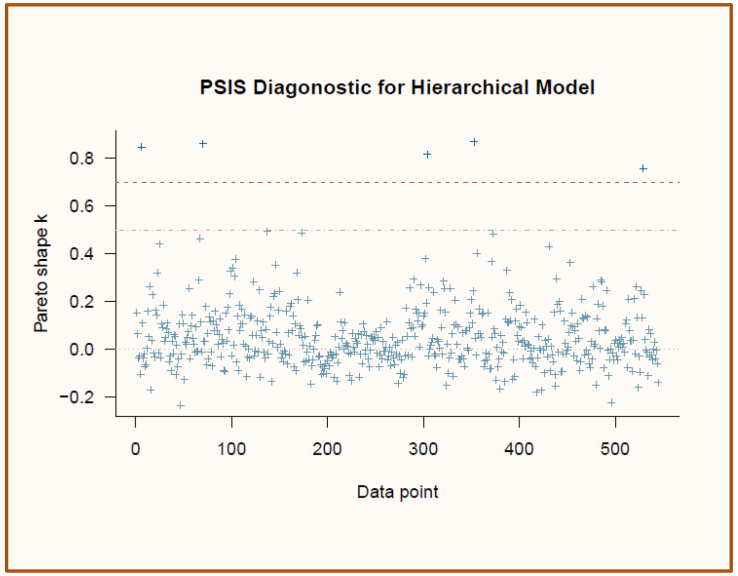
Gaussian Process

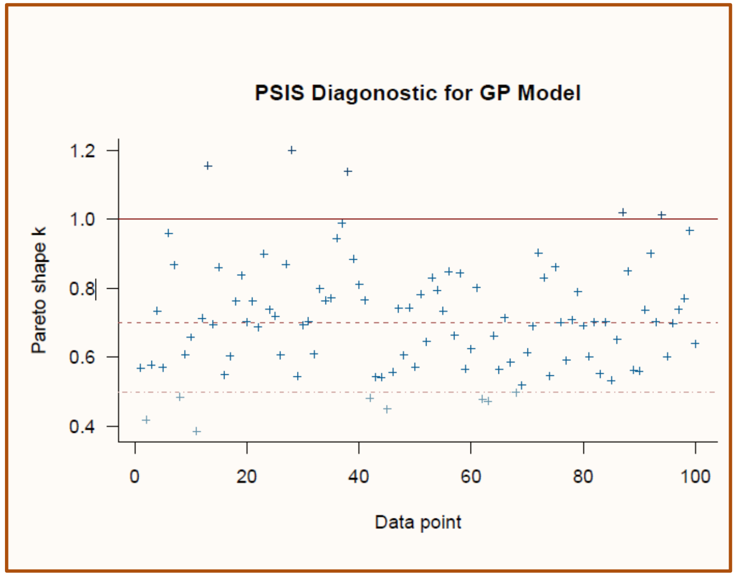
- Model $y \sim \mathcal{N}(f, \sigma^2)$ $f \sim GP(\mu(x), K(x|\theta))$
- where $K(x|\alpha,\rho,\sigma)_{i,j} = \alpha^2 exp(-\frac{1}{2\rho^2}\sum_{d=1}(x_{i,d} - x_{j,d})^2) + \delta_{i,j}\sigma^2$
- Stan model has been run for 4000 iterations, 4 Markov chains
- divided into train (100)and test (20) sets
- $\hat{R}$ = 1.003179
- n_{eff} > 0.01 for all parameters
- Bulk ESS over 100 for all the parameters
- No divergences in the GP



Model Comparison







- All Pareto k estimates are ok (k < 0.7)

	Estimate	SE
elpd_loo	-850.9	21.1
p_loo	12.5	1.5
looic	1701.8	42.1

- Good (99.1%), bad (0.9%)

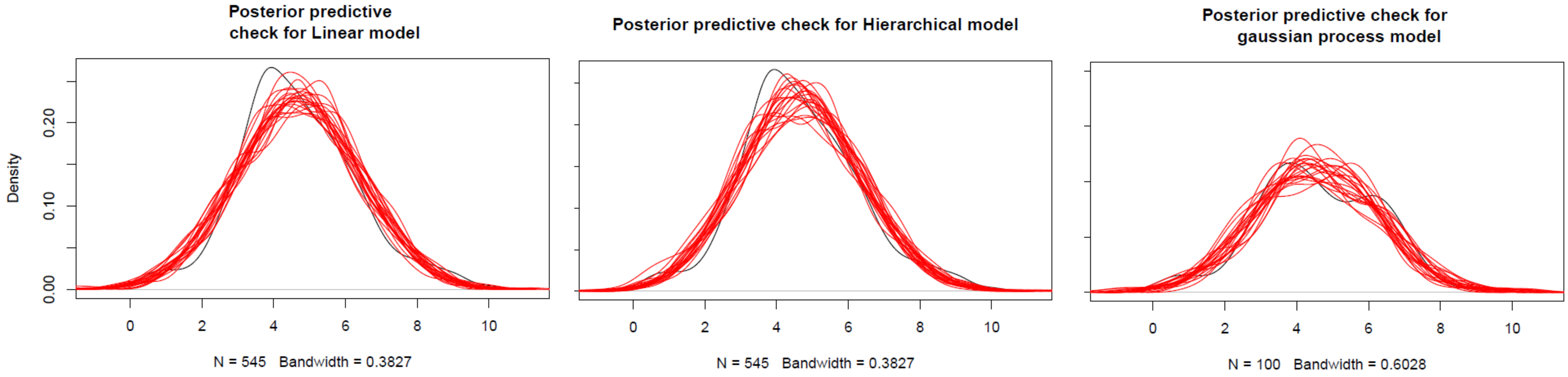
	Estimate	SE
elpd_loo	-871.4	22.0
p_loo	27.5	3.1
looic	1742.8	44.0

- Good (8.0%), ok (40.0%), bad (47.0%), v bad (5.0%)

	Estimate	SE
elpd_loo	-151.3	5.9
p_loo	63.1	4.6
looic	302.6	11.8

Posterior Predictive Checks

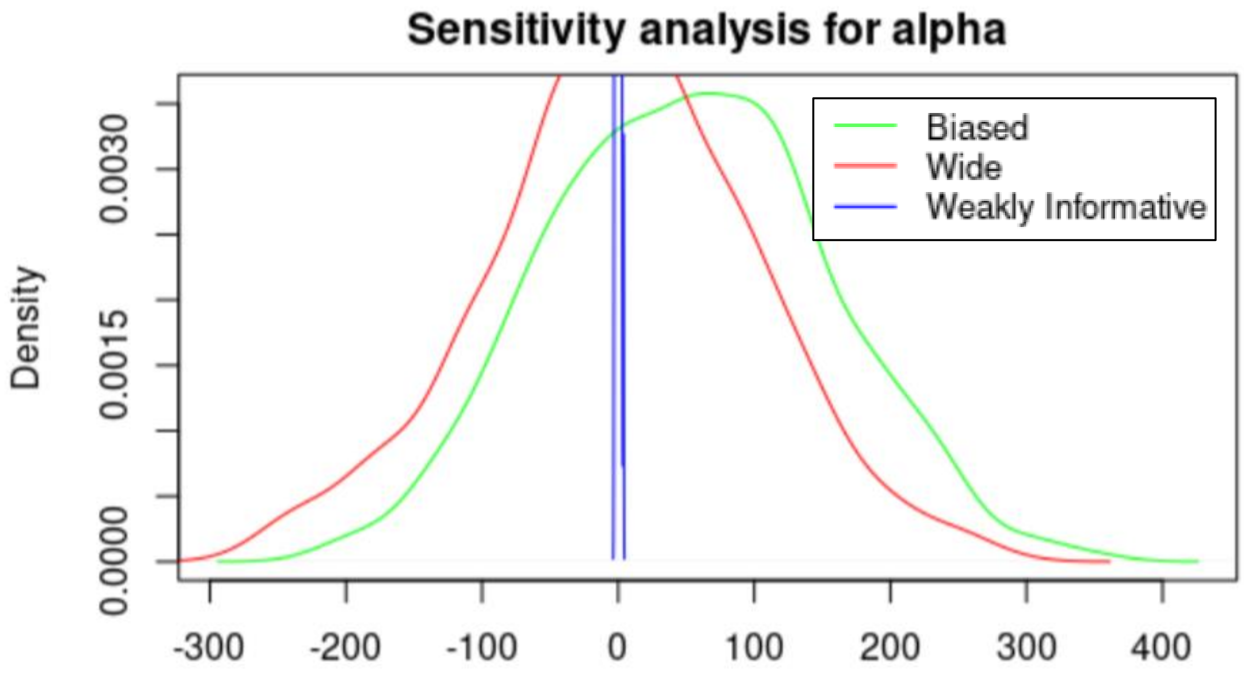
- Extracted the likelihood for "quantitative response" of the last 20 interactions
- Black → True value , Red → posterior predictive value



Sensitivity Analysis

❖ weakly informative priors ,wide priors and biased prior for linear model

Priors	Model	Pareto k estimates
alpha ~ normal(0,1) beta ~ normal(0,1) sigma ~ normal(0,100)	Linear	all k < 0.5
alpha ~ normal(0,100) beta ~ normal(0,100) sigma ~ normal(0,100)	Linear	all k < 0.5
alpha ~ normal(50,10) beta ~ normal(50,10) sigma ~ normal(0,100)	Linear	all k < 0.5



Sensitivity Analysis

- Weakly informative priors , wide priors and biased prior checks for

Hierarchical

Priors	Model	Pareto k estimates
mu ~ normal(0,1) tau ~ normal(0,1) sigma ~ normal(0,100)	Hierarchical	98.3% k < 0.5 0.7% –> 0.7 < k < 1
mu ~ normal(0,100) tau ~ normal(0,100) sigma ~ normal(0,10)	Hierarchical	97.6% k < 0.5 0.9% –> 0.7 < k < 1
mu ~ normal(50,10) tau ~ normal(50,10) sigma ~ normal(0,100)	Hierarchical	98.2% k < 0.5 1.3% –> 0.7 < k < 1

Gaussian

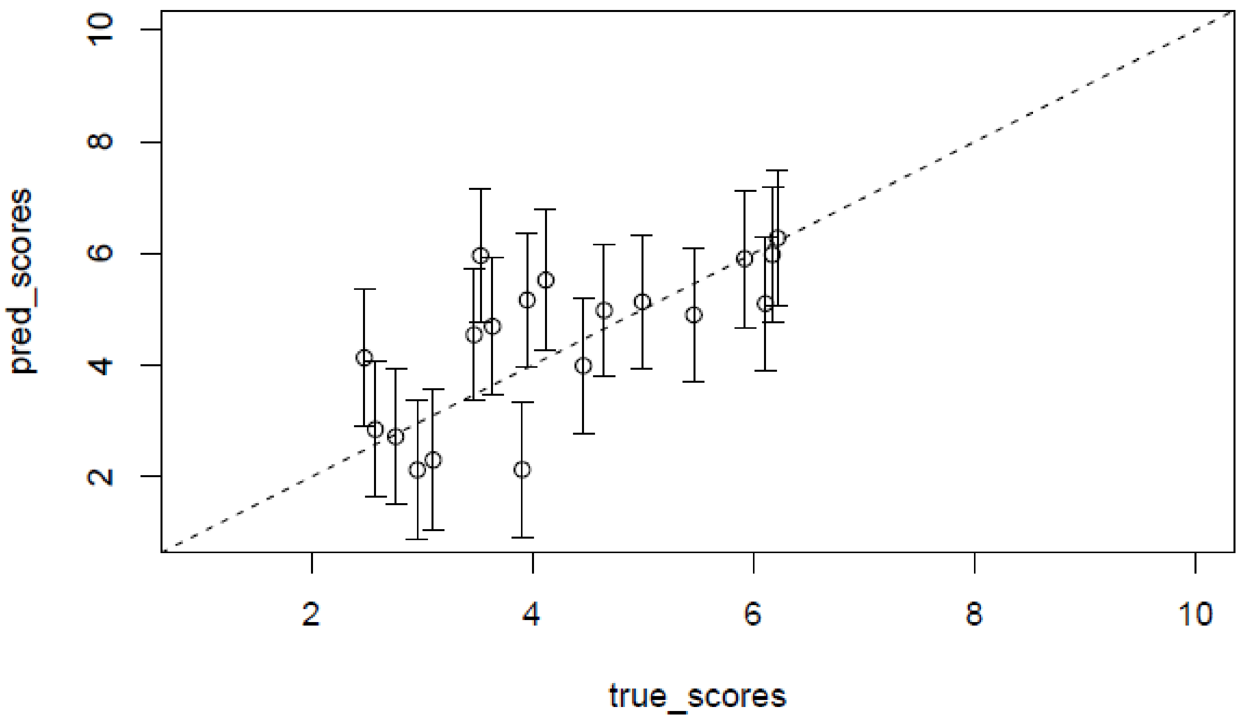
Priors	Model	Pareto k estimates
rho ~ inv_gamma(5,5) alpha ~ normal(0,1) sigma ~ normal(0,1) eta ~ normal(0,1)	Gaussian	9% k < 0.5 49% –> 0.7 < k < 1 7% –> k > 1
rho ~ inv_gamma(5,0.1) alpha ~ normal(0,100) sigma ~ normal(0,100) eta ~ normal(0,100)	Gaussian	0% k < 0.5 97% –> 0.7 < k < 1 3% –> k > 1
rho ~ inv_gamma(20,5) alpha ~ normal(50,10) sigma ~ normal(50,10) eta ~ normal(50,10)	Gaussian	0% k < 0.5 77% –> 0.7 < k < 1 21% –> k > 1

Predictive Performance Assessment

Metrics

- RSS = 24.69
- RMSE = 1.11
- Coefficient of Determination

  R2 = 0.45



Conclusion

- This report addressed the problem of predicting the toxicity of organic chemicals toward D. magna using QSAR dataset.
- RDCHI and MLOGP has the highest impact on Quantitative Response (Results from best model)

  With every 100% ↑ in RDCHI, there is 58 % ↑ in QR with 13% SD

  With every 100% ↑ in MLOGP, there is 44 % ↑ in QR with 6% SD

Applications [edit]

K_{ow} values are used, among others, to assess the environmental fate of persistent organic pollutants. Chemicals with high partition coefficients, for example, tend to accumulate in the fatty tissue of organisms (bioaccumulation).

https://en.wikipedia.org/wiki/Octanol-water_partition_coefficient

Aquatic Pollution ain't cool so don't be a fool!!!