# Statistical analysis of kinetic parameter data from the BRENDA database

This paper presents a range of multilevel Bayesian linear models describing kinetic parameter data from the online database BRENDA. We assess the models' qualitative fit to the available data and quantitatively compare their out-of-sample predictive performance. The best model is shown to compare favourably with the current state of the art, without needing to use costly external features such as chemical fingerprints. We discuss how the results of our analysis can be integrated into a kinetic modelling framework.

The data we used can be reviewed on Github and the analysis can be reproduced using XXX.

## Methods

## Statistical model

In order to aggregate the information about km parameters in the BRENDA database we use a multilevel Bayesian regression model with a nested structure. In this model the response variables are km values, which we assume are measured such that

$$\begin{aligned} y = \ln km &\sim student - t(4, \hat{y}, \sigma) \\ \hat{y}_{jk} &= \mu + \alpha_j^{EC3} + \alpha_{jk}^{EC4} \end{aligned}$$

where $\hat{y}_{jk}$ is the estimated true km value on natural logarithmic scale for an enzyme with EC3 number $j$ and ec4 number $k$. The symbols $\sigma$ and $\mu$ represent single unknown real numbers, $\alpha^{EC3}$ a vector of unknown real numbers and and $\alpha^{EC4}$ a matrix of unknown real numbers.

The prior distributions for $\alpha^{EC3}$ and $\alpha^{EC4}$ are as follows:

$$\begin{aligned} \alpha_j^{EC3} &\sim student\ t(0, sd^{EC3}) \\ \alpha_{jk}^{EC4} &\sim student\ t(0, sd_j^{EC4}) \end{aligned}$$

where $sd^{EC3}$ is an unknown positive real number and $sd^{EC4}$ is a vector of unknown real numbers (one for each EC3 number).

To complete our model we use informative priors for the remaining unknown model parameters:

$$\begin{aligned} \mu &\sim normal(0, 2) \\ \sigma &\sim lognormal() \\ sd^{EC3} &\sim lognormal() \\ sd^{EC4} &\sim lognormal() \end{aligned}$$

**Reasoning behind modelling choices**

The nested model design is motivated by our assessment of the available information. We judged that differences from the baseline are the km measurements are conditionally exchangeable given the same ec4 number, and that ec4 effects with the same ec3 number would be similar.

We used the student-t distribution with four degrees of freedom (t4) to model measurement errors and for our group-level regression models because the BRENDA data is known to contain many outliers. Since the t4 distribution has heavy tails compared to the normal distribution, we expected that this distribution would make the model less sensitive to outliers, resulting in better predictive performance. Testing confirmed this: models that used the t4 distribution resulted in better out-of-sample predictive accuracy than an equivalent models with normal-distirbuted errors [refer to appendix].

## Results

## Discussion