

Analysing enzyme databases with Stan and Arviz

Teddy Groves

- Introduction
- What we did
- Results
- Thoughts
- Thanks for listening!
- References

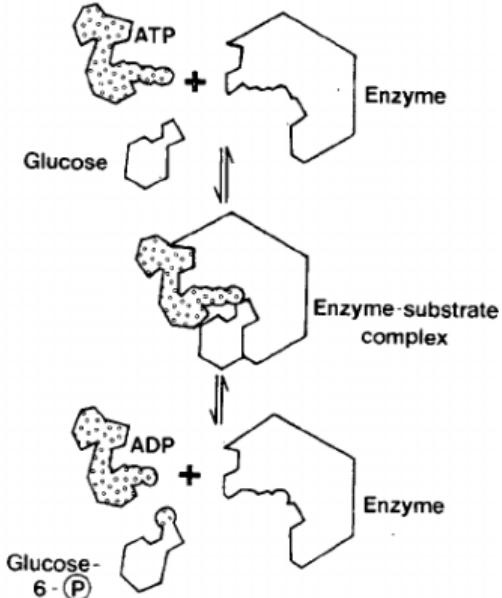
Introduction

Introduction

The good news

Since ~1920 we can parameterise enzyme-catalysed reactions:

$$v = E \cdot \frac{k_+^{cat} \frac{GLC}{k_{GLC}^M} \frac{ATP}{k_{ATP}^M} - k_-^{cat} \frac{G6P}{k_{G6P}^M} \frac{ADP}{k_{ADP}^M}}{(1 + \frac{GLC}{k_{GLC}^M})(1 + \frac{ATP}{k_{ATP}^M}) + (1 + \frac{G6P}{k_{G6P}^M})(1 + \frac{ADP}{k_{ADP}^M}) - 1}$$



Introduction

More good news

Kinetic parameters measurements are available online.

The screenshot shows the BRENDA website homepage. At the top, there is a navigation bar with a search field ('go to...'), a 'HOME' button, and a 'Classic view' link. To the right of the search field is the BRENDA logo, which features a stylized enzyme molecule icon followed by the word 'BRENDA' in large blue letters. Below the logo, the text 'The Comprehensive Enzyme Information System' is displayed. To the right of this text are links for 'Contact', 'UNI', 'Mehr', 'history', and 'all enzymes'. On the far right, there is a logo for 'Technische Universität Braunschweig'. Below the navigation bar is a search interface. It consists of a text input field containing 'Please enter a search term', a dropdown menu set to 'Enzyme, Ligand', a dropdown menu set to 'contains', and three buttons at the bottom labeled 'add search field', 'delete search field', and 'start search'.

Contribute to BRENDA! Your enzyme data is important for BRENDA. Send us your paper, and we will do all the work to include your data into our database. [More...](#)

Please enter a search term

Enzyme, Ligand contains

add search field delete search field start search

The bad news

- Disagreement between studies
- Relevant information can be inaccessible
- Many parameters have not been measured

Result: systems biology step zero

- ① Manually trawl the database
- ② Read the actual papers
- ③ Decide how to resolve disagreements

This is annoying!

Introduction Our plan

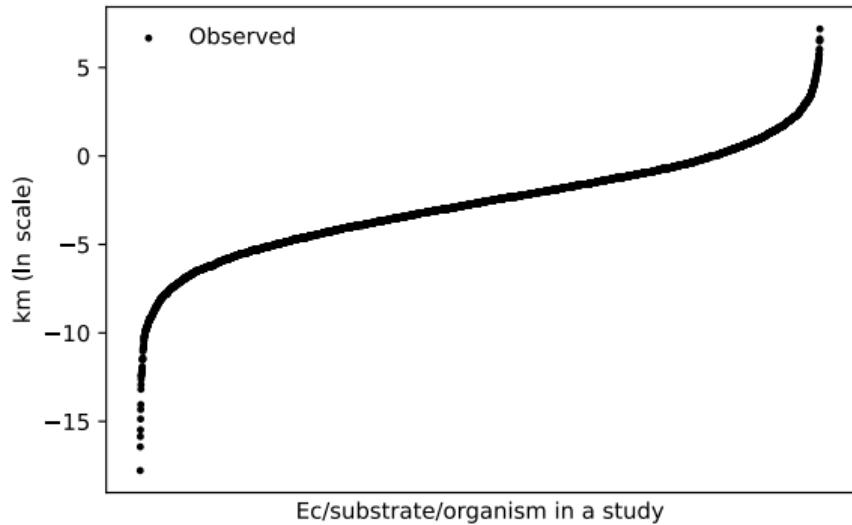
- ① Summarise the km information in BRENDA with a statistical model.
- ② Make the results easy to access

What we did

What we did

Fetch and process some data

- One 'observation' per study
- Three organisms (Human, yeast, E. coli)
- Natural enzyme/substrate combinations only



Try out some models

We settled on this measurement model from Borger, Liebermeister, and Klipp (2006):

$$\ln \hat{k}_m_{ijk} = \mu + \alpha_k^{sub} + \alpha_{jk}^{org:sub} + \alpha_{ik}^{ec:sub}$$

Write the model in Stan

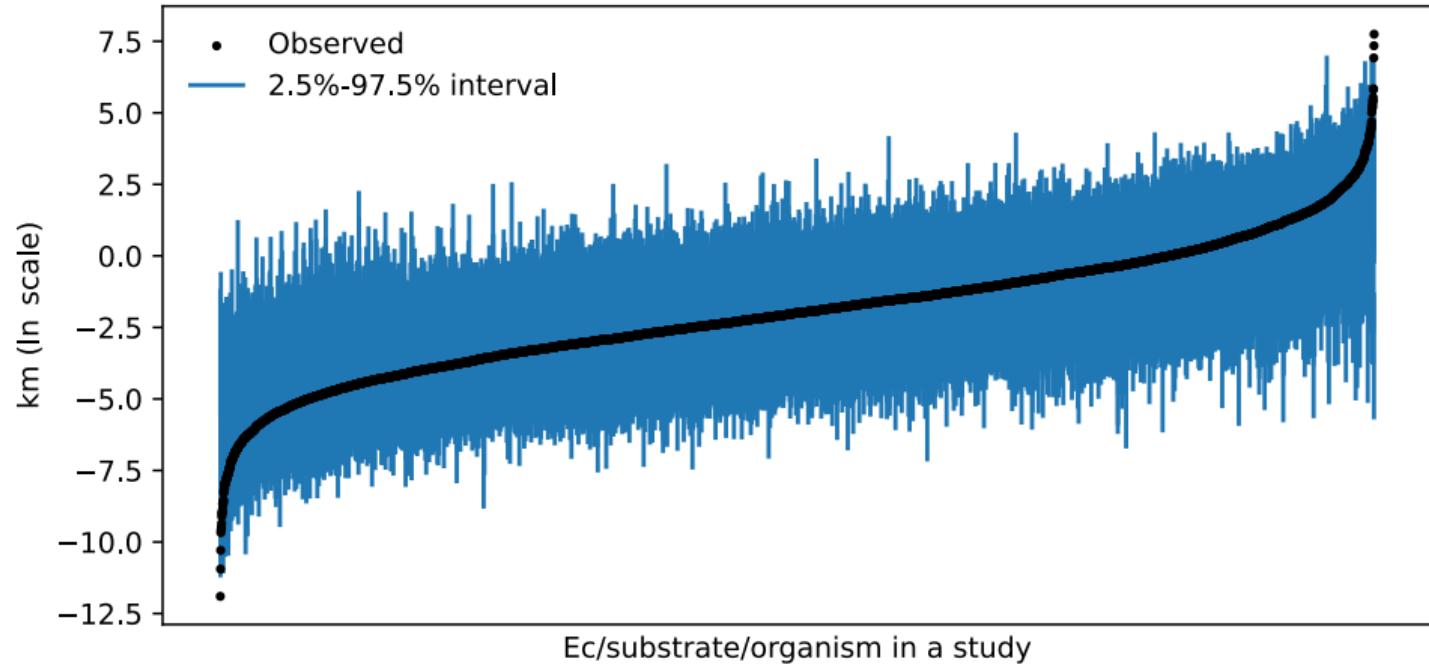
```
model {  
    if (likelihood){  
        y_train ~ student_t(nu, log_km[biology_train], sigma);  
    }  
    nu ~ gamma(2, 0.1);  
    sigma ~ normal(0, 2);  
    mu ~ normal(-2, 1);  
    a_sub ~ normal(0, tau_sub);  
    a_ec_sub ~ normal(0, tau_ec_sub);  
    a_org_sub ~ normal(0, tau_org_sub);  
    tau_org_sub ~ normal(0, 1);  
    tau_ec_sub ~ normal(0, 1);  
    tau_sub ~ normal(0, 1);  
}
```

Results

Results

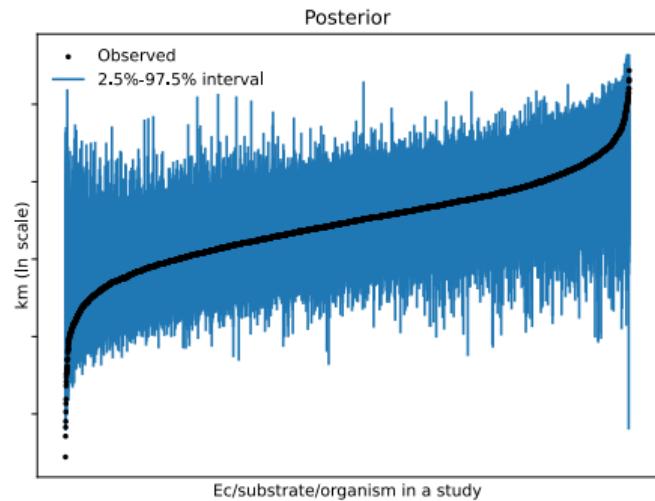
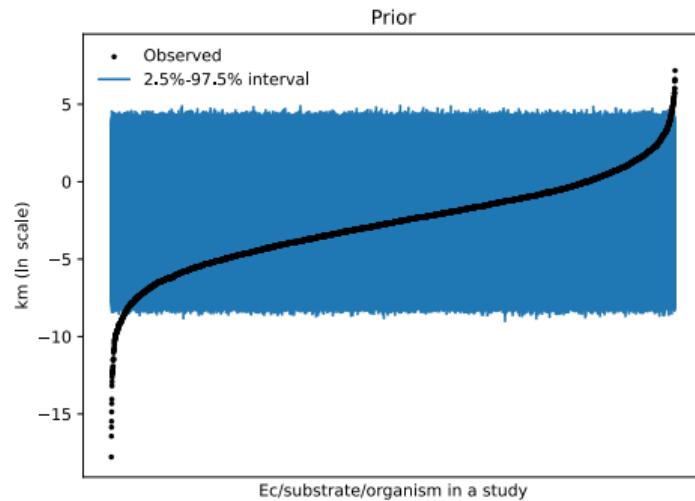
Fake data

Fake data posterior predictive distribution



Results

Real data



Results

Approximate cross-validation

	Estimate	SE
elpd_loo	-12549.28	86.78
p_loo	2681.82	-

There has been a warning during the calculation. Please check the results.

Pareto k diagnostic values:

		Count	Pct.
(-Inf, 0.5]	(good)	5914	92.6%
(0.5, 0.7]	(ok)	407	6.4%
(0.7, 1]	(bad)	65	1.0%
(1, Inf)	(very bad)	3	0.0%

Results

Exact cross-validation

```
: calculating out of sample log likelihoods for model blk...
:   split_4: -1477.2729990849998
:   split_3: -1481.00882687375
:   split_2: -1441.2448831225
:   split_5: -1477.1304594875
:   split_0: -1467.7910366725
:   split_7: -1480.53716923375
:   split_9: -1504.9884546787503
:   split_8: -1502.4060516825002
:   split_6: -1483.4119273775
:   split_1: -1515.77739092
: total out of sample log likelihood: -14831.569199133752
```

Results

Webapp

Choose which marginal posterior distribution you'd like to see!

Organism

Escherichia coli

EC4

1.1.1.169

Substrate

2-dehydropantoate

[Download a csv table of model results.](#)

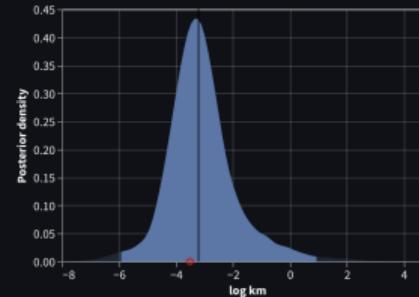
[Download a csv table of measurements.](#)

See [here](#) for the model's source code and for instructions to reproduce the full posterior distribution.

What does BRENDA say?

This webapp shows marginal posterior distributions for Km parameters from a model we trained on data from the [BRENDA database](#).

The graph shows the posterior distribution (blue) and experimental measurements (red).



Marginal posterior summary:

Mean: -3.0908038874325

Standard deviation: 1.23

1% quantile: -5.99

Median: -3.20807

99% quantile: 1.0

Below is a table of all the measurements.

	reference	km	log km
0	687675	0.03	-3.506558

Thoughts

Thoughts

- This data is hard to model
- File-based cmdstanpy workflow is great!
- llik + yrep = long io

Thanks for listening!

Thanks for listening!

Thanks for listening!

Thanks for listening!

Our emails:

- tedgro AT biosustain.dtu.dk
- aretsi AT biosustain.dtu.dk

The project on github: -https://github.com/biosustain/brenda_km

novo nordisk fonden

References

References I

Borger, Simon, Wolfram Liebermeister, and Edda Klipp. 2006. “Prediction of Enzyme Kinetic Parameters Based on Statistical Learning.” *Genome Informatics* 17 (1): 80–87.