FAIR Bayesian Optimal Experimental Design in Systems Biology

Sebastian Höpfl[1], Jürgen Pleiss[2], Jan Range[2] and Nicole Radde[1]

[1]University of Stuttgart, Institute for Systems Theory and Automatic Control, Pfaffenwaldring 9, 70569 Stuttgart, Germany
{sebastian.hoepfl, nicole.radde}@ist.uni-stuttgart.de
[2]University of Stuttgart, Institute of Biochemistry and Technical Biochemistry, Allmandring 31, 70569 Stuttgart, Germany

The acquisition of experimental data in Systems Biology is time-consuming and costly. On the one hand, this is due to expensive chemicals, such as antibodies, on the other hand, due to the need for extensive preparation, incubation, and purification procedures of experiments. Finding the best experiment design helps to get the maximum possible information out of limited experimental data, thereby minimizing cost and time. Here, we present first results of Bayesian Optimal Experiment Design (BOED) in a Findable, Accessible, Interoperable, and Reusable (FAIR) fashion. In particular, our approach makes use of the Systems Biology Markup Language (SBML), the de facto modeling standard in Systems Biology, and is set up in an easily applicable and reusable way. Models, data, and code will be stored on Fairdomhub.

BOED was applied to Michaelis-Menten enzyme kinetics with different initial substrate concentrations c_{init} (Figure 1). In the first attempt, three different experimental designs were investigated with artificial data. The first design includes three measurements with initial substrate concentrations larger than K_M, the second design includes three measurements with initial substrate concentrations lower than K_M, and the third design includes only two measurements but with lower and higher initial substrate concentrations than K_M. Our results show that in the first design only k_{kat} but not K_M can be identified. In the second design, k_{kat} and K_M are correlated. Only in the third design with only two experiments, both parameters can be identified and the IG was highest. These results demonstrate that proper planning of the experimental design can improve the estimation of the parameters and at the same time reduce the number of experiments needed. Now, the next step is to predict the optimal starting concentrations via BOED.

In practice, we want to use the Kullback-Leibler divergence (KLD) to calculate the difference between the prior and the posterior. The design that maximizes the KLD has thereby the largest information gain. Mutual information is the expected KLD between the prior ($p(\theta)$) and the posterior $p(\theta|d, y)$, with design d and data y:

$$U(d) = \int_{\Theta} \int_{Y} p(\theta, y|d)[\log p(\theta, y|d) - \log p(y|d) - \log p(\theta)]dyd\theta \qquad (1)$$

Both, the posterior $P(\theta|d, y)$ and the evidence $p(y|d)$ are needed for the calculation of the mutual information utility. Therefore, we want to apply the dynesty [1] algorithm. Dynesty is a dynamic nested sampling algorithm that estimates the posterior and evidence simultaneously and is already implemented in the pyPESTO [2] toolbox which works with SBML. Nested sampling splits the likelihood slices and samples them individually via n_{live} live samples. The

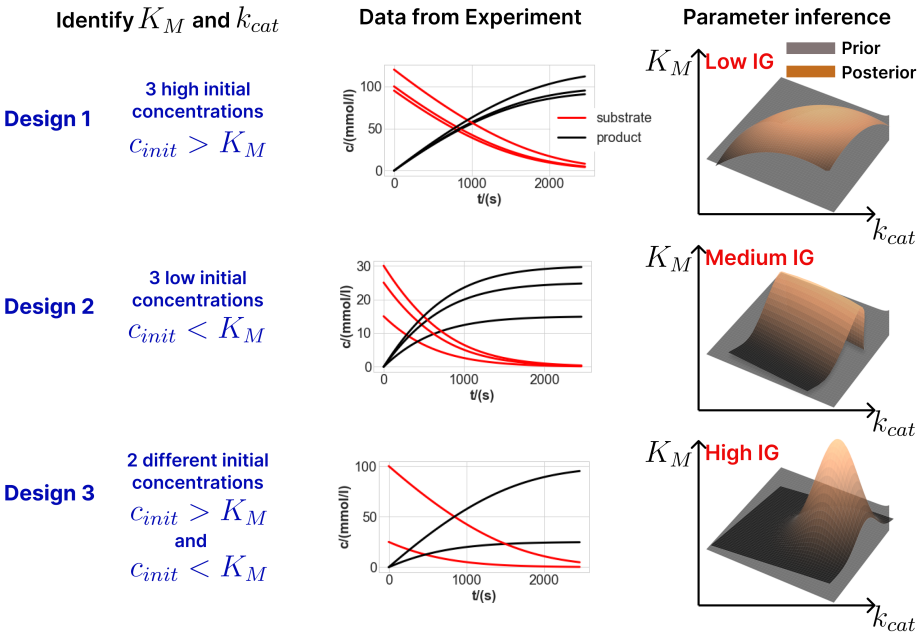

Figure 1: **Bayesian Optimal Experiment Design applied to enzyme-driven substrate conversion for the identification of the parameters K_M and k_{cat}.** BOED maximizes the information gain between the prior and posterior. Three experimental designs, with (1) initial concentrations higher than K_M, (2) initial concentrations lower than K_M and (3) a combination of higher and lower concentrations are considered. Their respective metabolization time curves are used to estimate the posterior (bronze) for each design. The Information Gain (IG) quantifies the difference between Prior and Posterior.

weighted sum of samples of these likelihood slices approximates the evidence. Dynamic nested sampling adapts the number of live points in the sampling process dynamically, to get a more accurate estimate of the evidence.

In conclusion, we plan to use dynamic nested sampling for BOED in a FAIR way. Therefore, we will use toolboxes that are maintained by the Systems Biology community and are applicable to SBML. This will be done for an artificial Michalis-Menten model first and afterward applied to real enzymatic data.

References

[1] Speagle, Joshua S. (2020): dynesty: a dynamic nested sampling package for estimating Bayesian posteriors and evidences. In: *Mon. Not. R. Astron. Soc.* 493 (3), S. 3132–3158. DOI: 10.1093/mnras/staa278 .

[2] Schälte, Yannik; Fröhlich, Fabian; Stapor, Paul; Vanhoefer, Jakob; Weindl, Daniel; Jost, Paul Jonas et al. (2022): pyPESTO - Parameter EStimation TOolbox for python: Zenodo.

Funding

Funded by the Research Unit Programme FOR 5151 QuaLiPerF (Quantifying Liver Perfusion–Function Relationship in Complex Resection—A Systems Medicine Approach) by grant no. 436883643 and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2075 - 390740016. We acknowledge the support by the Stuttgart Center for Simulation Science (SimTech).