# Small Dataset Acquisition for Machine Learning Analysis of Industrial Processes with Possible Uncertainties

**Xukuan Xu,** Felix Conrad, Andreas Gronbach, Michael Möckel
Technische Hochschule Aschaffenburg

Email: Xukuan.Xu@th-ab.de

# About us

Xukuan Xu earned a Diplom degree in mechanical engineering from Technische Universität Dresden, Germany in 2020. He is currently pursuing a doctoral degree at Technische Hochschule Aschaffenburg, specializing in the field of small batch production with a focus on machine learning, process monitoring, and anomaly detection.

TH Aschaffenburg

TU Dresden

# BMBF Competence Cluster InZePro
# Project KIproBatt

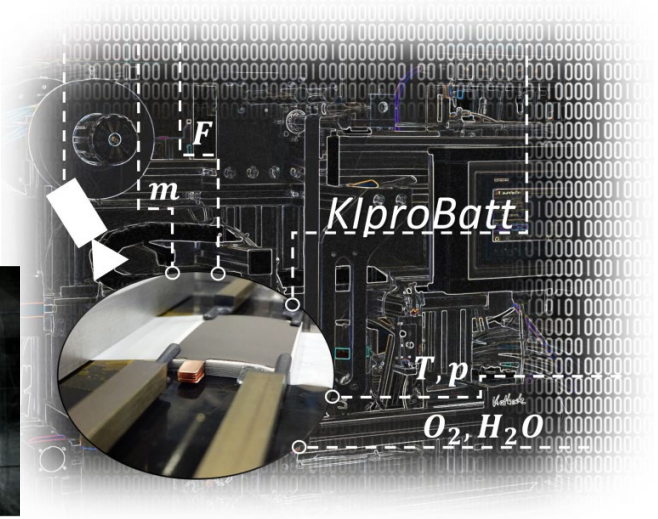
TH Aschaffenburg
university of applied sciences

KIproBatt:
Intelligent Battery Cell Production with AI-based Process Monitoring based on a Generic System Architecture

Goal:
- Process modeling ---- Hybrid AI
- Representation of process correlations ---- Data-driven approach
- Process monitoring ---- Sensor hardware & dataspace

Specification:
- An intelligent battery cell production
- A production-oriented generic system architecture
- General applicability & transferability

Homepage: https://kiprobatt.de/wiki/Main_Page

# Aims and contributions of our paper

TH Aschaffenburg
university of applied sciences

- Data preparation is the bottleneck of ML applications according to a survey from Crowdflower [2]
- This problem is often overlooked. In most cases, the datasets are unthinkingly pre-existing
- The quality of the dataset determines the upper limit of data analysis

In our paper, we :

1) discussed the characteristics of small-data context with process uncertainties based on practical examples
2) developed an adapted design of experiments (DOE) aiming at preparing datasets for ML applications

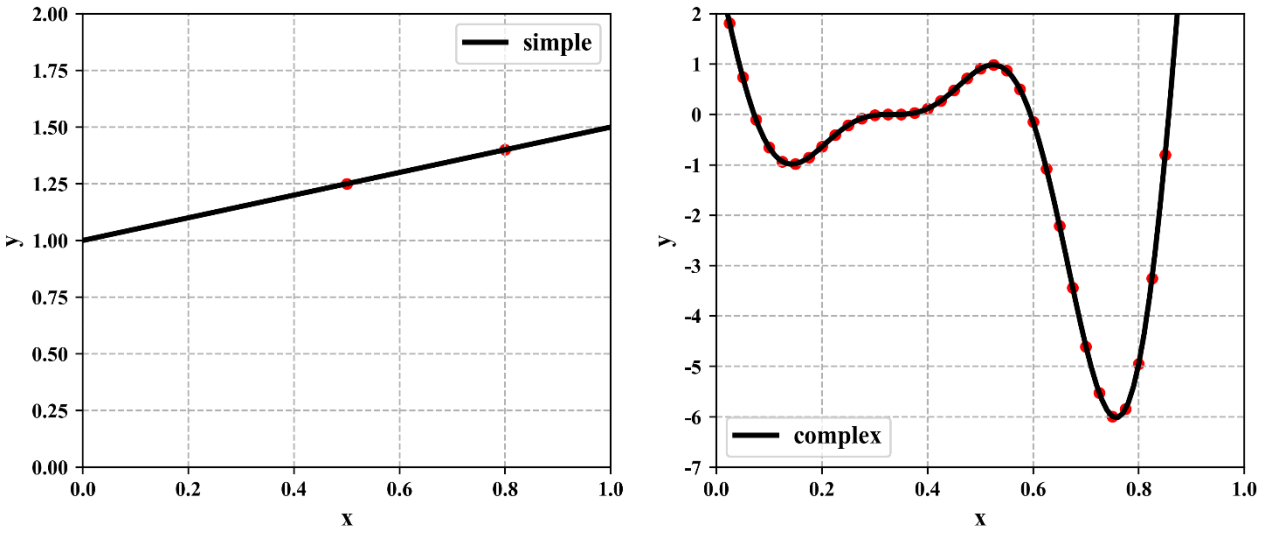
# a. Small-data problem

Under which circumstances will small-data problem appear:

- **Small-batch production** is often unavoidable in laboratory research, on a pilot production stage prior to upscaling, or in customer-specific (individualized) manufacturing [5]
- Data acquisition is limited by **budget or time constraints** to datasets with <1000 elements

However, the amount of data is not necessarily dominant factor

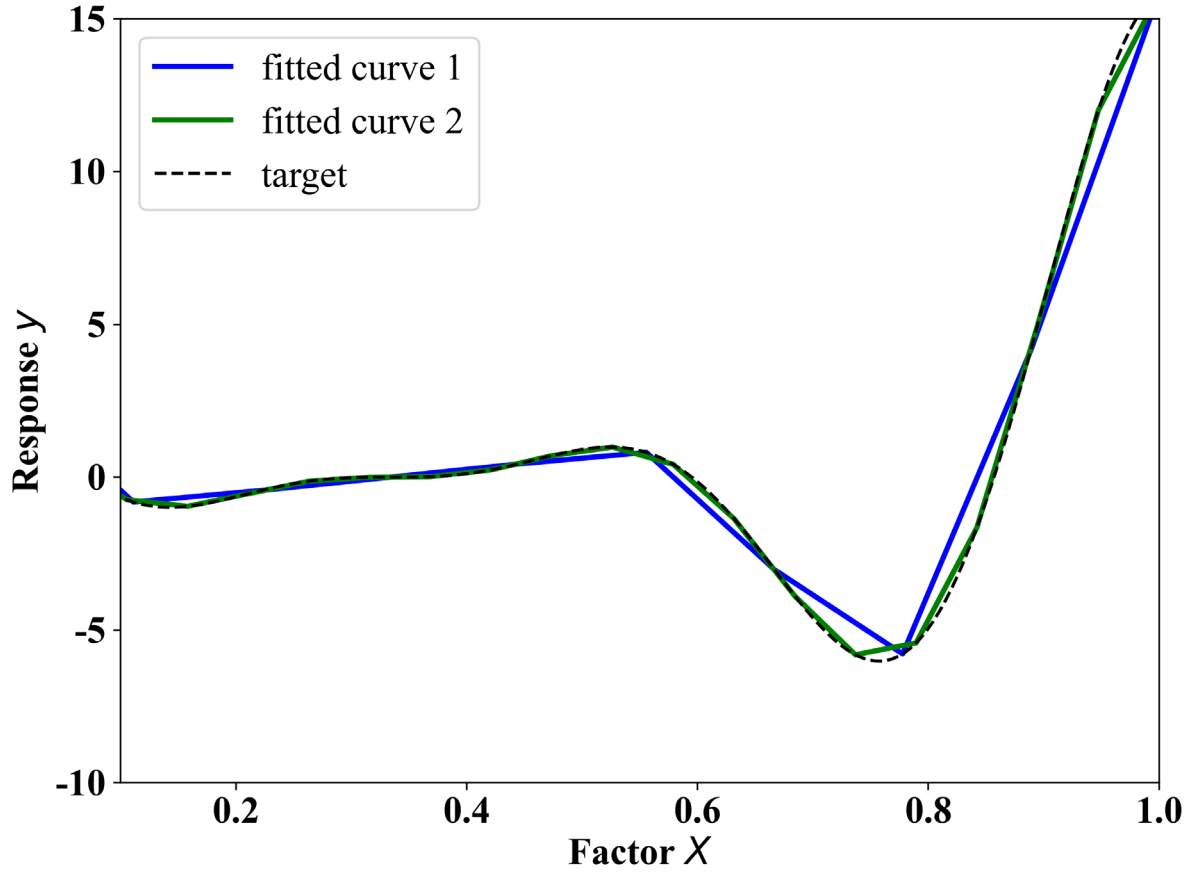
- The number of required data depends on the **complexity** of the process
- The particular **distribution** of selected data points affects the outcomes of analysis



# a. Small-data problem

the fundamental characteristics of small-data context:

- under-sampling of the parameter space
- a lack of convergence of the ML models.

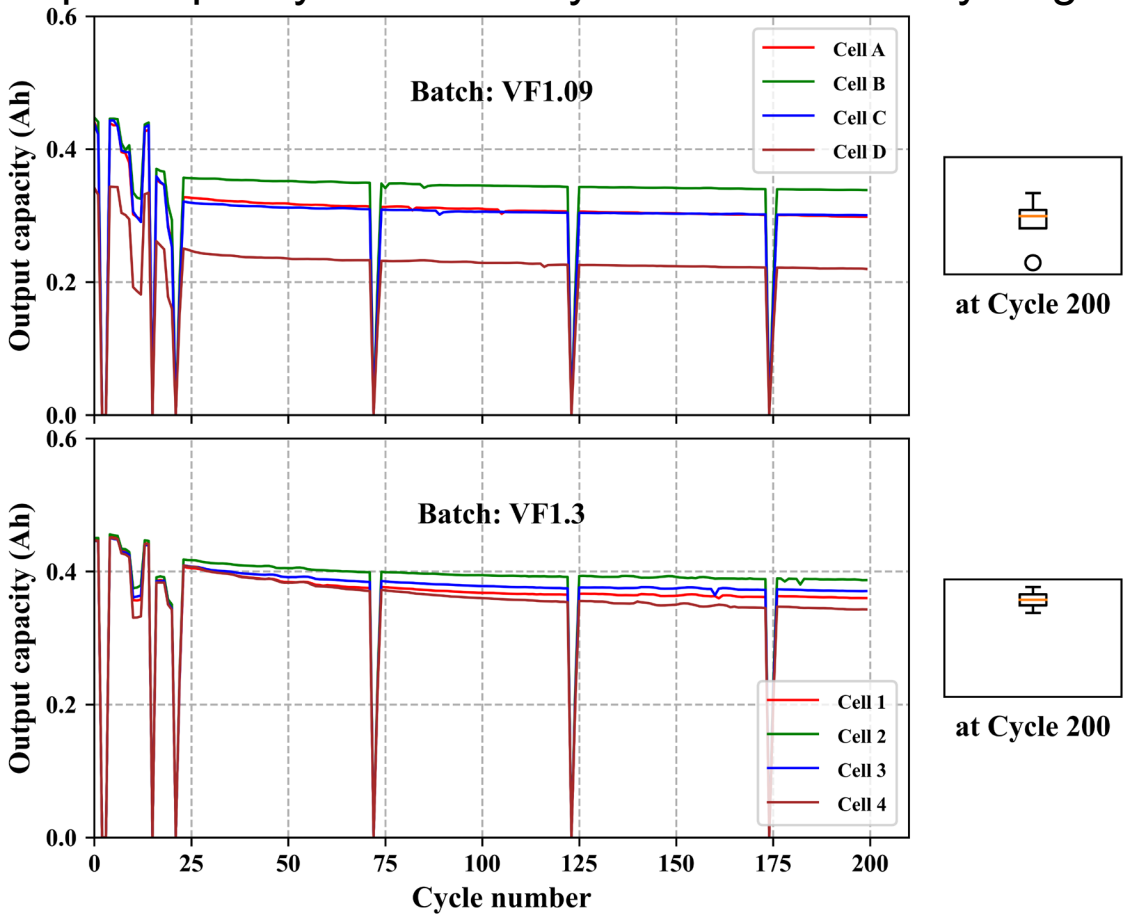


# b. Process uncertainties

**Process uncertainty:**
a statistical spreading in the target responses

- measurement uncertainties
- unavoidable fluctuations of the process parameters
- some potential variables uncontrolled
- human factor

Cell output capacity related to cycle number in a cycling test



# c. Existing DOE strategies

TH Aschaffenburg
university of applied sciences

According to DEAN et al. [3], the two main purposes of DOE are:

I. to find the control variables that give rise to an **optimal** response
II. to obtain **a mathematical description** for the pre-defined dataspace in order to predict further responses

Traditional DOE workflow:

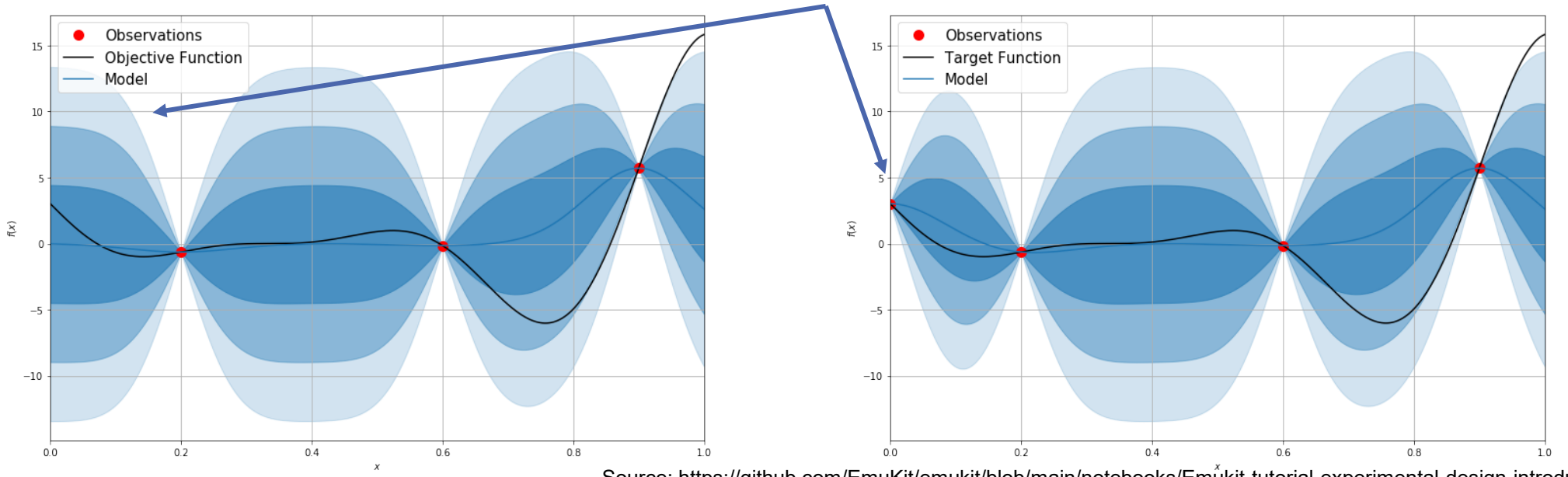
1. Determine the factors of interest and the response
2. Identify the significant factors
   - (fractional) factorial design: full factorial design, Plackett–Burman design
3. Establish a regression model
   - central composite design, Box-Behnken design

Drawbacks:
Inflated data volume with multiple factors
outdated analysis techniques ( in comparison with the development of ML )
Predetermined, non-adaptive

# d. Adaptive DOE strategy

Iterative data acquisition schemes have been discussed, e.g., Emukit [9] provides such a model-based iterative DOE scheme within a Bayesian optimization framework.
3 steps to generate data points:

- fit a prediction model to the existing data (initial dataset)
- find the next point with the highest marginal predictive variance as predicted by the prediction model
- add this new data point to the existing dataset

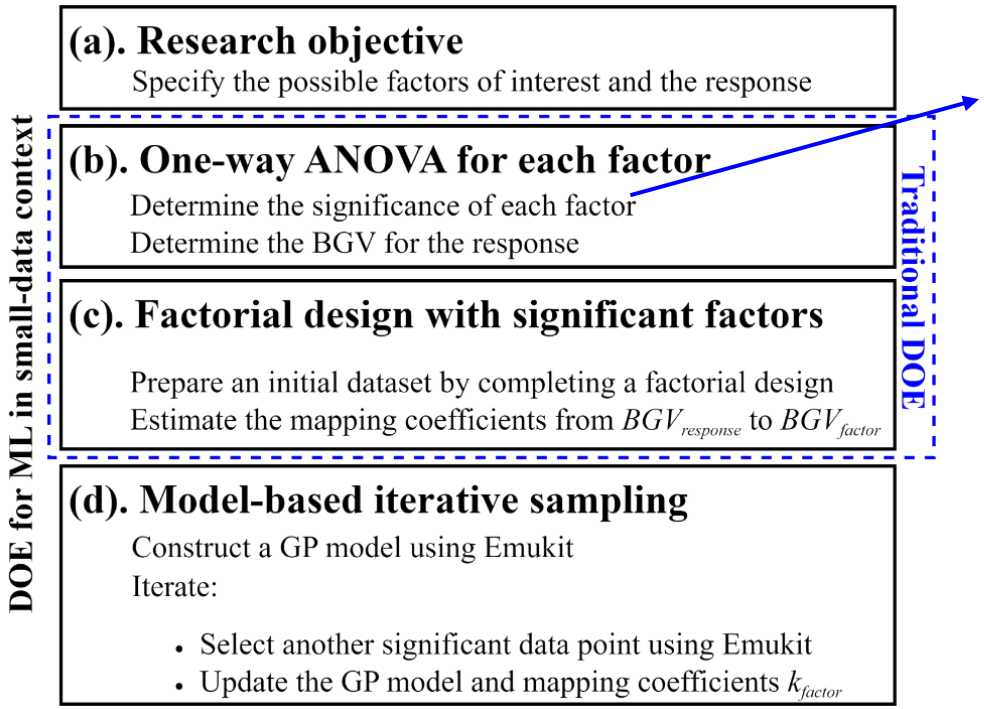



Source: https://github.com/EmuKit/emukit/blob/main/notebooks/Emukit-tutorial-experimental-design-introduction.ipynb

# e. Our proposed DOE strategy for small-data context

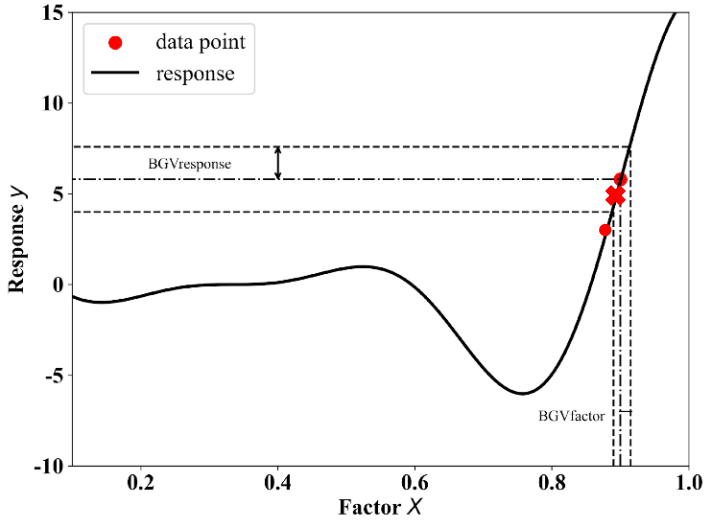
TH Aschaffenburg
university of applied sciences

Proposed DOE workflow in small-data context




BGV: between-group variance
GP model: Gaussian Process model

Assuming that the levels we set for the one factor are too **close** to each other, then the statistical spreading due to within variance may limit the **distinguishability**.
In other words, once the P-value exceeds the predefined level of significance ($p > p_0$), the considered factor should be judged as insignificant within this interval.
If this principle is applied to select the next data point, it can determine whether this data point is valid for the considered factor.

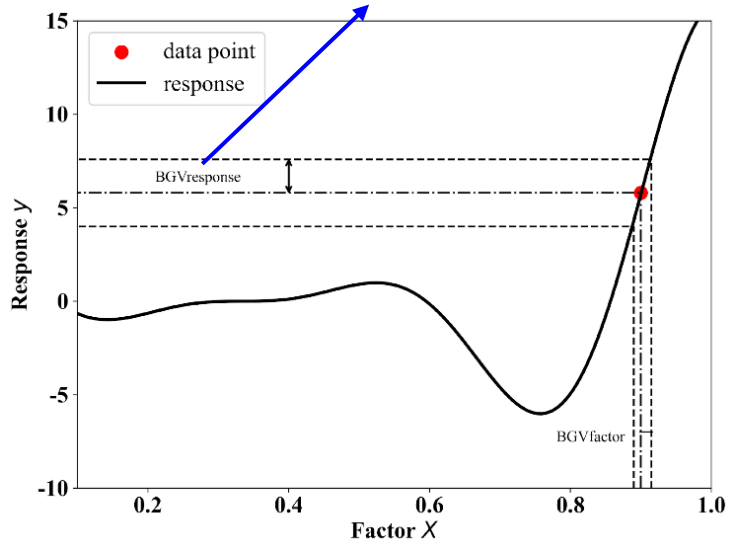


# e. Our proposed DOE strategy for small-data context

ANOVA: OC versus EV

| Source | DF | Adj SS | Adj MS | F-Value | P-Value |
|---|---|---|---|---|---|
| Electrolyte | 1 | 0.011542 | 0.011542 | 8.16 | 0.029 |
| Error | 6 | 0.008482 | 0.001414 | | |
| Total | 7 | 0.020024 | | | |

$$BGV_{response,min} = \sqrt{F\ value_{1,6,a=0.05} * Adj\ MSE}$$




Next to the **Electrolyte Volume** (EV), experts believe [6] that **Drying Time** (DT), **Wetting Time** (WT), **Coating Defects** (CD) on electrodes, and **Stacking Accuracy** (SA) also have impact on the **Output Capacity** (OC):

$$OC_{cycle\ 200}(X_{EV}, X_{DT}, X_{WT}, X_{CD}, X_{SA})$$

$$k_{factor} = \frac{\partial OC_{cycle\ 200}}{\partial X_{factor}}$$

For instance:

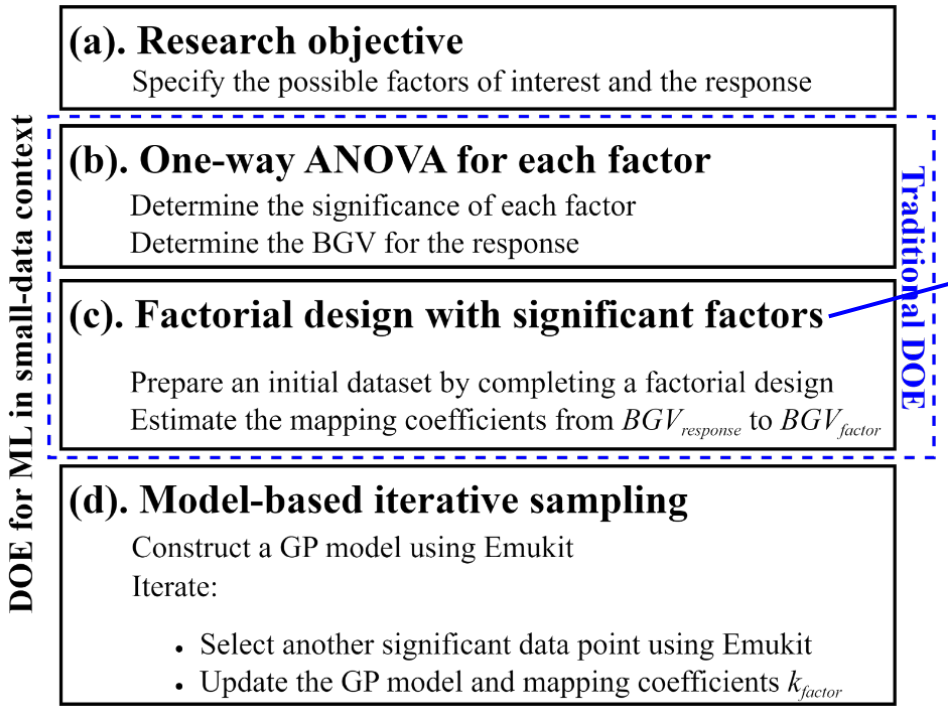
$$OC_{cycle\ 200} = \sum_{i=1}^{5} k_i X_i + b$$

$$BGV_{factor,min} = \frac{BGV_{response,min}}{k_{factor}}$$

# e. Our proposed DOE strategy for small-data context

TH Aschaffenburg
university of applied sciences

Proposed DOE workflow in small-data context




Next to the **Electrolyte Volume** (EV), experts believe [6] that **Drying Time** (DT), **Wetting Time** (WT), **Coating Defects** (CD) on electrodes, and **Stacking Accuracy** (SA) also have impact on the **Output Capacity** (OC):

$$OC_{cycle\ 200}(X_{EV}, X_{DT}, X_{WT}, X_{CD}, X_{SA})$$

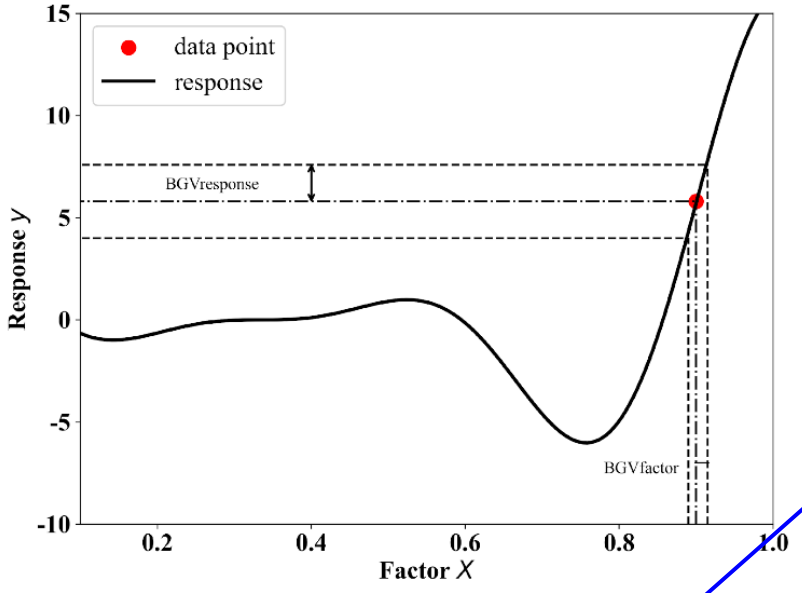
$$k_{factor} = \frac{\partial OC_{cycle\ 200}}{\partial X_{factor}}$$

For instance:

$$OC_{cycle\ 200} = \sum_{i=1}^{5} k_i X_i + b$$

$$BGV_{factor,min} = \frac{BGV_{response,min}}{k_{factor}}$$

# e. Our proposed DOE strategy for small-data context




- $BGV_{factor,min}$ defines an environment around each factor value where no significant data points can be chosen
- Under this framework, we can proceed the iterative sampling in step (d) until all data points for a ML dataset have been collected
- Each new data point will be used to update the model and the mapping coefficients to determine a more accurate estimate for $BGV_{factor,min}$

# Future work

TH Aschaffenburg
university of applied sciences

- We will conduct a comprehensive analysis of our proposed DOE workflow with a FEM-simulation:
  - Collecting data by running the simulation with limited number of access
  - Add noise function to emulate process uncertainties
  - Comparing different sampling strategies by investigating the correlation between the sampling strategy used and it´s corresponding ML model's performance on the test data

- Our DOE strategy will be used to guide the production of cells in the KIproBatt program

# Reference


[1] F. Conrad, M. Mälzer, M. Schwarzenberger, H. Wiemer, and S. Ihlenfeldt, "Benchmarking AutoML for regression tasks on small tabular data in materials design", Sci Rep, vol. 12, no. 1, Art. no. 1, pp. 19350, Nov. 2022, doi: 10.1038/s41598-022-23327-1.

[2] Figure Eight. *CrowdFlower: Data science report*. [Online]. Available from: https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf [retrieved: 02, 2023].

[3] A. Dean, D. Voss and D. Draguljić, Design and Analysis of Expermients, 2nd Edition. New York, NY: Springer, 2017.

[4] X. Xu et al., *KIproBatt: Exploring smart battery cell production based on a generic system architecture and an AI-enhanced process monitoring*. [Online]. Available from: https://doi.org/10.13140/RG.2.2.11573.76006 2021.11.07

[5] J. Fleischer, G. Lanza and K. Peter, "Quantified Interdependencies between Lean Methods and Production Figures in the Small Series Production," Manufacturing Systems and Technologies for the New Frontier, pp. 89–92, 2008, doi: 10.1007/978-1-84800-267-8_17.

# Reference


[6] M. Westermeier, *Qualitätsorientierte Analyse komplexer Prozessketten am Beispiel der Herstellung von Batteriezellen.* [online]. Available from: https://www.mec.ed.tum.de/fileadmin/w00cbp/iwb/Institut/Dissertationen/322_Westermeier_Markus.pdf [retrieved: 02, 2023].

[7] R.A. Fisher, The Arrangement of Field Experiments in Breakthroughs in Statistics. New York, NY: Springer, 1992.

[8] L. Salmaso et al., "Design of experiments and machine learning to improve robustness of predictive maintenance with application to a real case study", Communications in Statistics - Simulation and Computation, vol. 51, no. 2, pp. 570–582, Feb. 2022, doi: 10.1080/03610918.2019.1656740.

[9] A. Paleyes et al., "Emulation of physical processes with Emukit". arXiv, Oct. 25, 2021. doi: 10.48550/arXiv.2110.13293.

[10] M. Zhang, A. Parnell, D. Brabazon, and A. Benavoli, "Bayesian Optimisation for Sequential Experimental Design with Applications in Additive Manufacturing". arXiv, Nov. 23, 2021. doi: 10.48550/arXiv.2107.12809.

[11] Z. Liu et al., "Machine learning with knowledge constraints for process optimization of open-air perovskite solar cell manufacturing", Joule, vol. 6, no. 4, pp. 834–849, Apr. 2022, doi: 10.1016/j.joule.2022.03.003.