**Data Science Challenges in Computational Chemistry**

M. Sc. Julia Jasper  
M. Sc. Nicolas Tielker  
M. Sc. Yannic Alber  
Prof. Dr. Stefan M. Kast  

_Fakultät für Chemie und Chemische Biologie_  
_Physikalische Chemie III_
Data & Science in Chemistry

Domain Expertise

Data Science

Mathematics

Computer Science

Statistical Research

Data Processing

Machine Learning

Chemists are (usually) in the danger zone.
Simulation-based methods

- Application starts out with available structural data
- Propagating atomic positions for a set of molecules yields highly correlated individual data points
- Millions of snapshots (results have to be extracted from trajectory)
- Tools for analysis often originate from statistics (Markov chain models, correlation analysis)
Data structures

- Distribution of solvent around a solute
- Volumetric data on 3D grids (density or energy value at each point)
- Large files → storage problem
- Volumetric data collapsed to scalars
- Both data types can be used for machine learning
Compensating for physical approximations

- Clean data but errors caused by physical approximations
- Complexity unaccounted for by high level data fitted through regression models
- Expensive and slow (larger systems can be unfeasible due to exponential scaling)
Regression Models on Synthetic Data

Practical applications of Data Science

- Correcting errors within the data generated using physics-based methods by training low level models
- Using physics to compensate for insufficient experimental data sources
- Correlate formally disjoint observables by fitting to independent quantities
- Retrieving information and knowledge from fitted models
- *Grand challenge:* **Be fast and predictive**

Further challenges

- Research data management: which types of data have to be stored and which can be discarded?
- NFDI (Nationale Forschungsdateninfrastruktur)
- Integration of Data Science in curricula