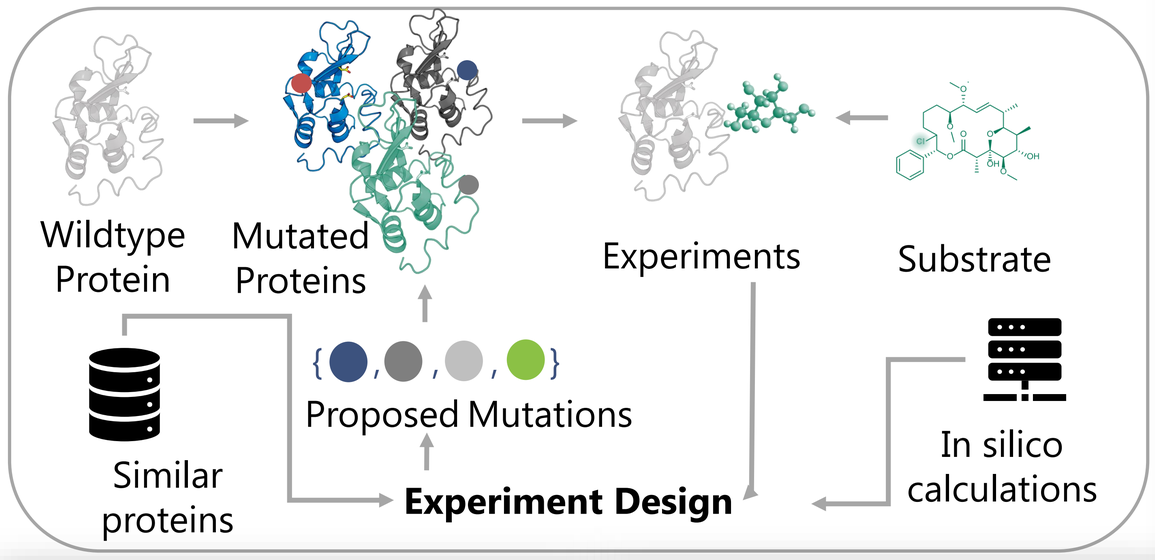
Modern Sequential Experiment Design for Protein Discovery

Mojmír Mutný, Andreas Krause

Experiment Design Paradigm (Active Learning)



➢Challenges

- ➢Large combinatorial space
- ➢Complicated interaction model
- ➢Experimental noise

➢Novelties

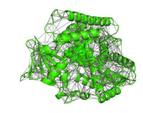
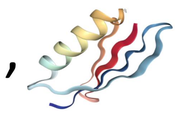
- ➢High-throughput assay
- ➢Novel Machine Learning
- ➢Novel Experiment Design Approaches
- ➢Novel Computational Methods

Model Fitting – understanding what matters

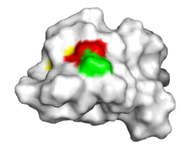
➢ **Ingredients:**

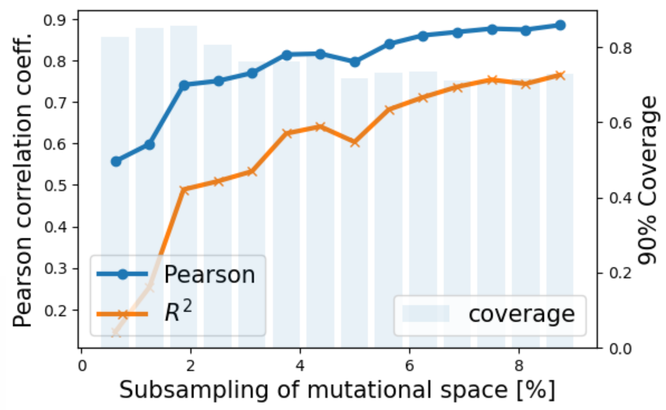
Key ingredient: similarity notion

- ➢ Metric learning/Kernels
- ➢ Neural network embeddings
- ➢ Self-supervised learning

k( , )

Can we identify similarity form the data?





GB1 dataset

- ➢ Antibody binding protein
- ➢ 20^4 variants screened
- ➢ Gaussian process model
- ➢ Learning kernel on sample
- ➢ With less than 2% search space we have a useful model
- ➢ Despite prediction being poor – the uncertainty is well-calibrated

Typical Pipeline

Mutant approximate structure

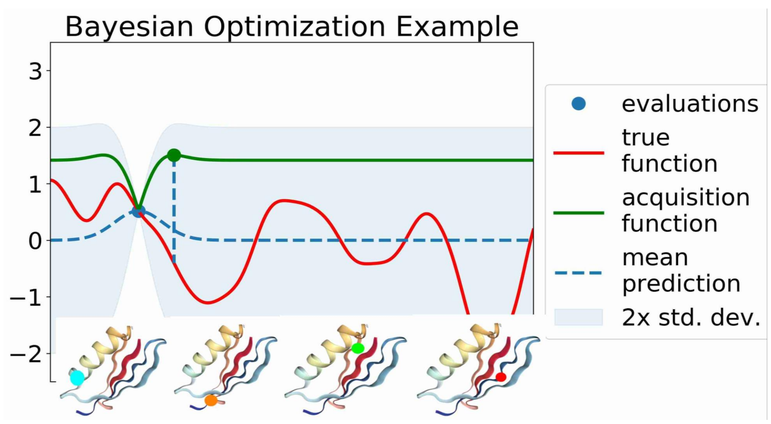
- **Rosetta Descriptor**
- **Geometric Descriptors**

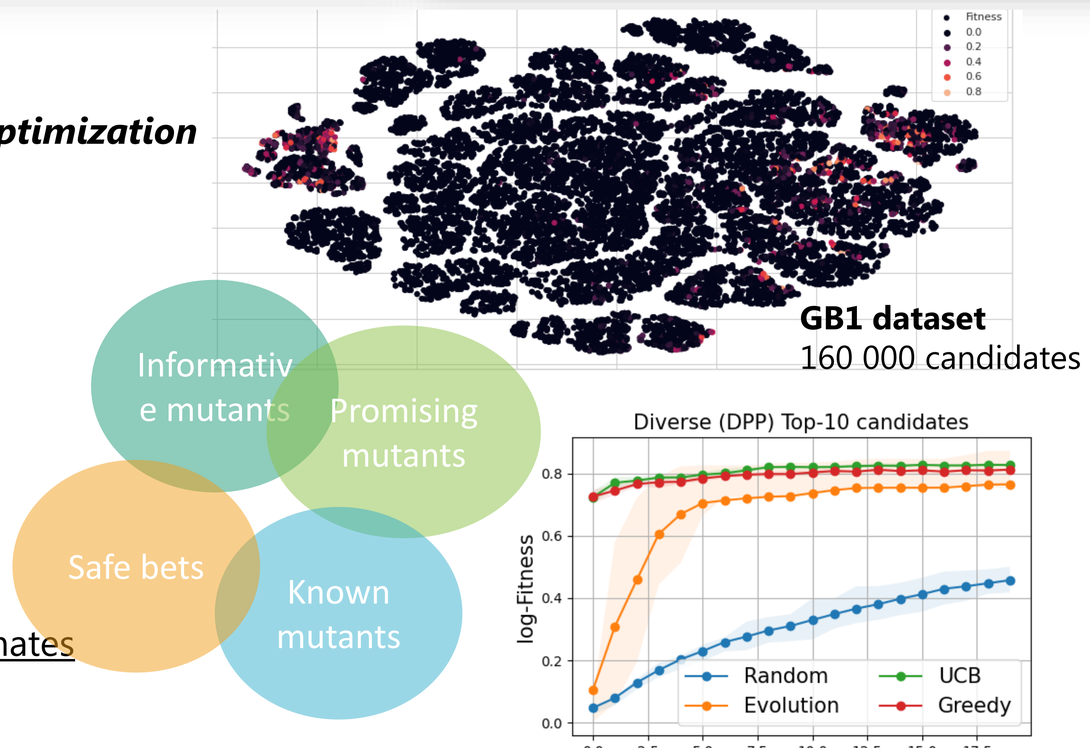
Mutant sequence

- **Chemical Descriptors**

Uncertainty and Applications

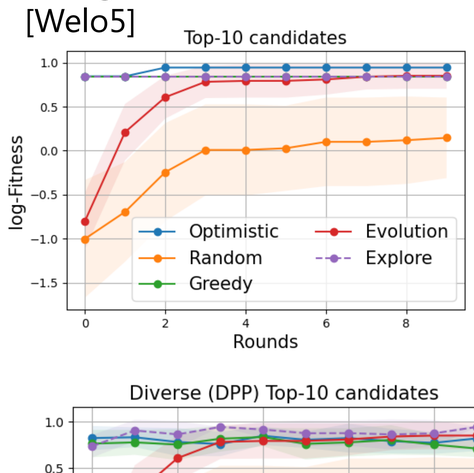
Experiment Design to Optimize = Bayesian Optimization

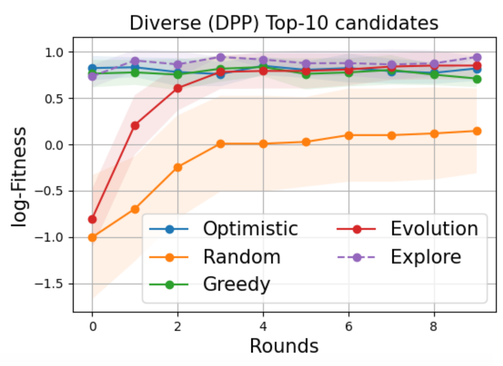




Halogenase dataset

[Welo5]





Diverse batch selection

- ➢ Screening for variants should be diverse – *What is diverse?*
- ➢ The similarity notion implies diversity as well
- ➢ Mathematical model: *Determinantal point processes -> sampling*

Future Outlook

- Scaling to larger combinatorial spaces beyond 20^5
- Use of advanced models for similarity
- Multi-fidelity – i.e. using simulators to understand structure
- Multi-objectivity

References

- Nature Communications v. 13, Art. no.: 371 (2022)
- Nature Communications v. 11, Art. no.: 1782 (2020)
- Nature Communications v. 13, Art. no.: 3788 (2022)
- PMLR 151:7031-7054, (2022).
- PNAS 110 (3) E193-E201 (2013)
- Proceedings of the IEEE Volume: 104, 1, 148-175 (2016)
- https://doi.org/10.48550/arXiv.1705.0060