TransMEP: Transfer Learning for **Mutation Effect Prediction**

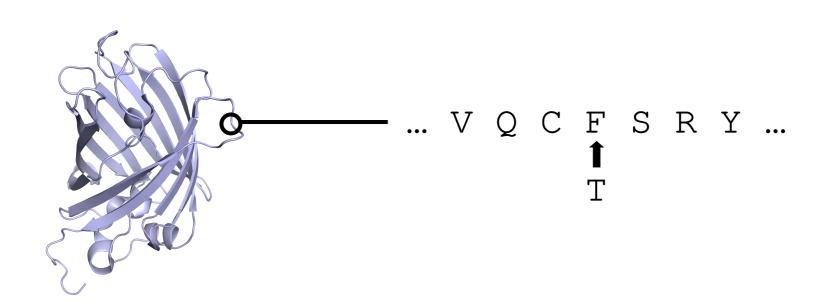


An Easy-to-Use Pipeline

Tilman Hoffbauer¹, Birgit Strodel^{1,2}

- ¹Computational Biochemistry Group, Institute of Biological Information Processing (IBI-7: Structural Biochemistry), Forschungszentrum Jülich, Germany
- ² Computational Biochemistry Group, Institure of Theoretical and Computational Chemistry, Heinreich Heine University Düsseldorf, Germany

Protein Engineering demands Mutant Assessment



- Mutations change protein properties, e.g. activity
- Mutant assessment: Quantify mutation effects
- Aritificial mutations are used in protein engineering

Example: GFP (UniProt ID 5N9O)

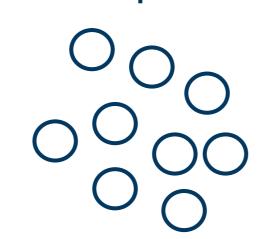
Shortcut: Machine Learning (ML) Promising Knowledge Candidates Training Sample Model

- ML is cheaper & faster than experimental assessment
- Promising candidates require experimental validation
- But: Naive ML is not accurate enough [1, 2]

Variant Prediction using Transfer Learning Pipeline

250 million unlabelled **Embed variants Evolutionary Scale** Task-specific Embedded space of interest Modelling [3] model protein sequences Model Protein (Gaussian Language Process) Model

Intuition: Embedded space provides a better separation of the mutants by interesting values, e.g. activity and structure.

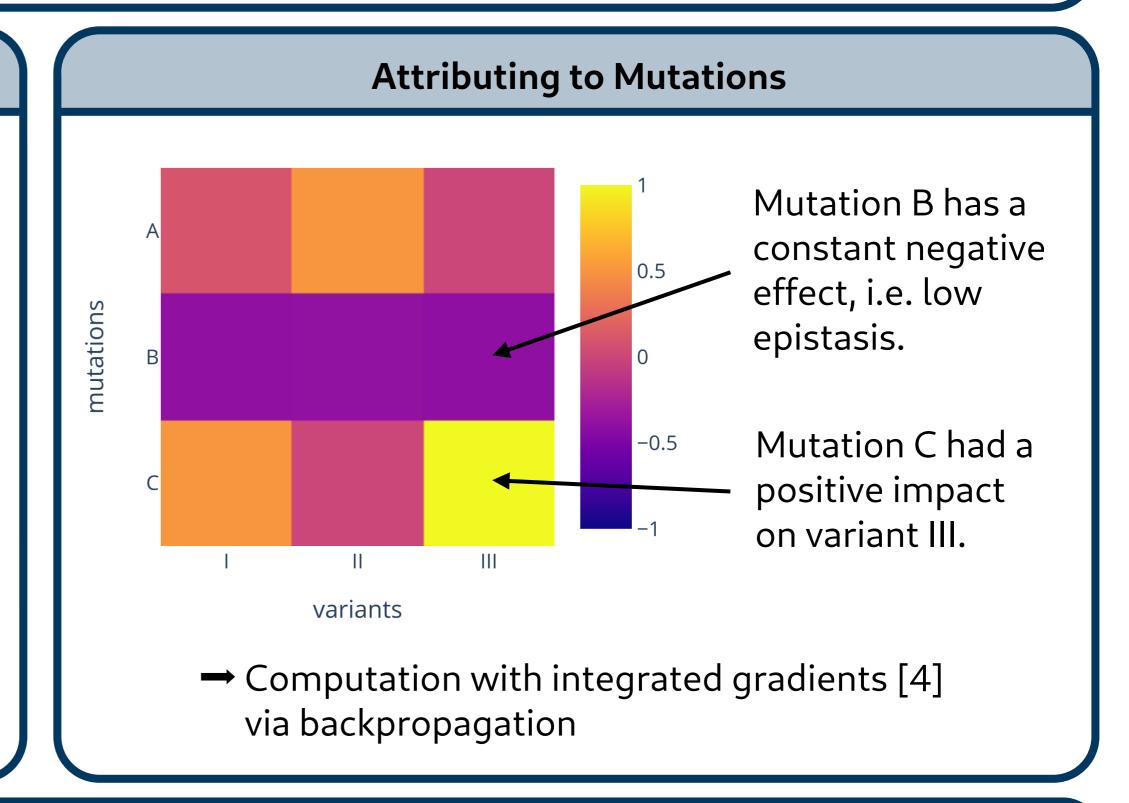


This pipeline was implemented in an easy-to-use program which is available as a webpage (via Google Colab) and command line interface.

Training variant V had a high influence 0.6 on the prediction of variant III. Training variant VI 0.2 was unimportant for

→ Can be inferred from model weights

Attributing to Training Samples



TransMEP: An Easy-to-Use Pipeline

- Improved mutation effect prediction using transfer learning
- Simple usage, no installation required

variants

- Allows attribution of predictions to individual mutations
- Detailed reports for analysis

Publication will follow soon! E-Mail: tilman.hoffbauer@rwth-aachen.de We thank Anna Jäckering for fruitful discussions and RWTH Aachen University for HPC resources.

training variants

these predictions.