WEIHENSTEPHAN · TRIESDORF
University of Applied Sciences

SynBiofoundry @TUM

TUMCS

TUM

Advancing Protein Engineering with Large Language Models

Prof. Dr. Dominik Grimm

Weihenstephan-Triesdorf University of Applied Sciences

TUM Campus Straubing for Biotechnology and Sustainability

SynBioFoundry, TUM Campus Straubing for Biotechnology and Sustainability

Technical University of Munich, TUM School of Computation, Information and Technology

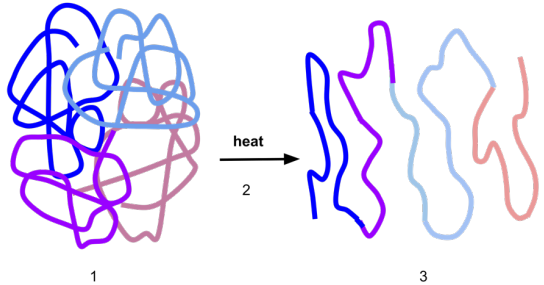
GrimmLab

Accurate Prediction of Thermophilic and Mesophilic Proteins

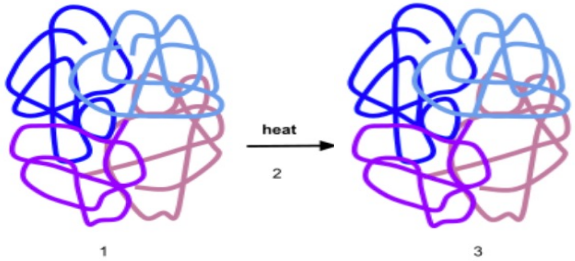
Thermostability of Proteins

The **thermostability of proteins** is an essential **property** that is important in many biotechnological fields, such as **enzyme engineering** and **protein-hybrid optoelectronics**

Example: High-power light emitting diodes have working device temperatures above 70°C



https://en.wikipedia.org/wiki/Thermostability#/media/File:Process_of_Denaturation.svg

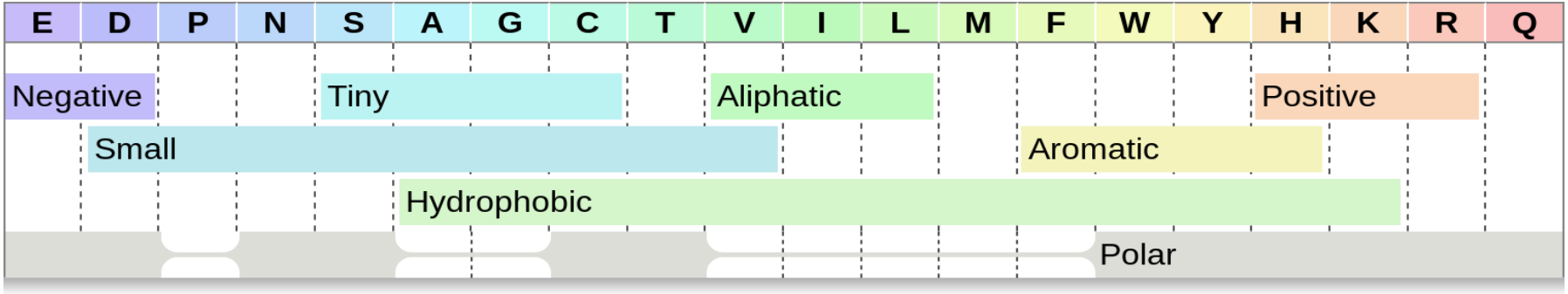


→ It is essential to accurately identify thermostable proteins

Problem: Enormous search space of potential candidates

Machine learning can be used to predict whether a protein is thermophilic or mesophilic

Physicochemical Properties as Features



https://commons.wikimedia.org/wiki/File:Rainbow_boxes_displaying_the_properties_of_amino_acids.png

- Derive physicochemical properties for each amino-acid in a protein sequence as features:
  - **Basic descriptors**, such as weight, charge, polarity, mean cdW volume etc..
  - **Residue composition**
  - **Physicochemical properties**, such as composition and distribution
- Train **classical discriminative machine learning** models on thermophilic and mesophilic protein sequences (e.g. Zhang and Fang 2007; Lin and Chen 2011; Charoenkwn et al. 2021; Ahmed et al. 2022)

Data

- We derived **data from previously published** studies (e.g. Zhang and Fang, 2007; Lin and Chen, 2011; Ahmed et al. 2022) and **cleaned up the dataset**, e.g. removed duplicated and overlapping sequences, merged them with the latest UniPort entries etc..
- In addition, we **collected new data** using different resources and databases, e.g. TEMPURA (Sato et al., 2020)
- **Removed evolutionarily related sequences** with a similarity of more than 40%
- Derived **599 physicochemical features**

Full dataset

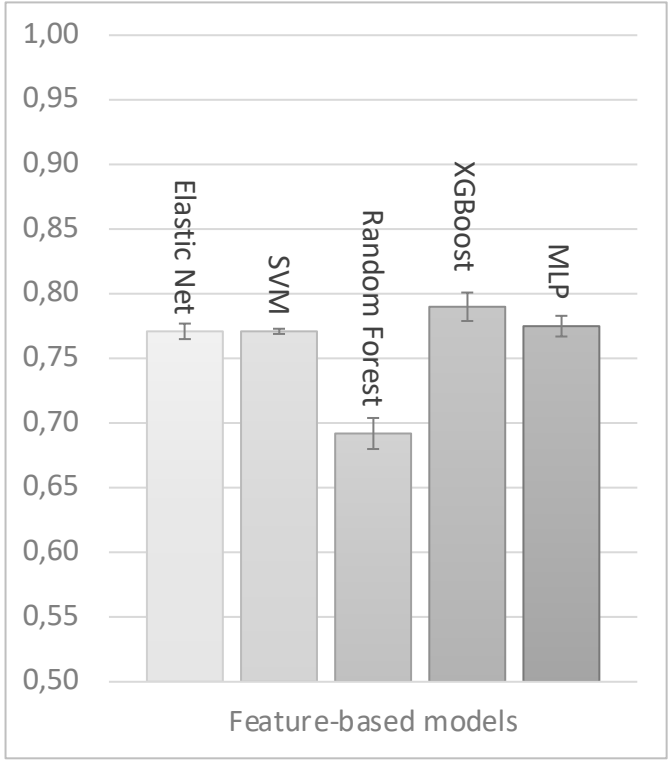
Class	Sequences
non-thermophilic	4545
thermophilic	2864

Cleaned and filtered dataset

Class	Sequences
non-thermophilic	3440
thermophilic	1699

Nested cross-validation with **Bayesian hyperparameter optimization**

Matthew's Correlation Coefficient (MCC) on test data in nested cross-validation



$$MCC = \frac{tn \cdot tp - fn \cdot fp}{\sqrt{(tp + fp)(tp + fn)(tn + fp)(tn + fn)}}$$

- +1 best agreement between predicted and actual values
- 0 no agreement
- -1 perfect misclassification
- **Measurement is unaffected by unbalanced class ratios**

Similarity between human language and protein sequences

There is a similarity between human languages and protein sequences

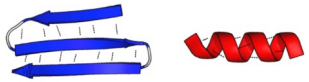
Letters: A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

Words: DNA | protein | is | language | human

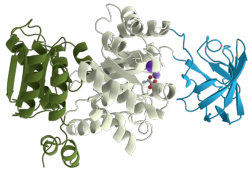
Sentences: A language model is a probability distribution over sequences of words

Amino Acids: A R N D C E Q G H I L K M F P S T W Y V

Secondary Structure

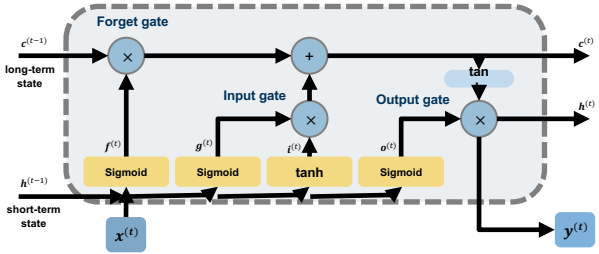


Domains

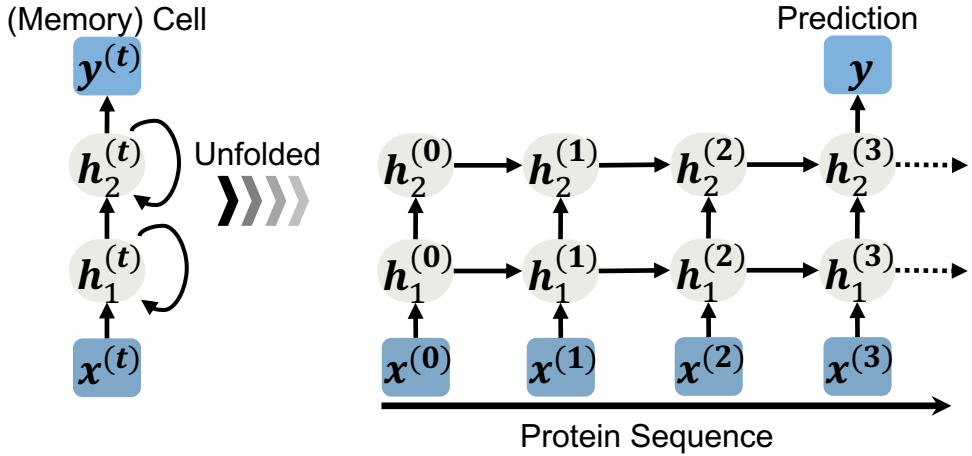


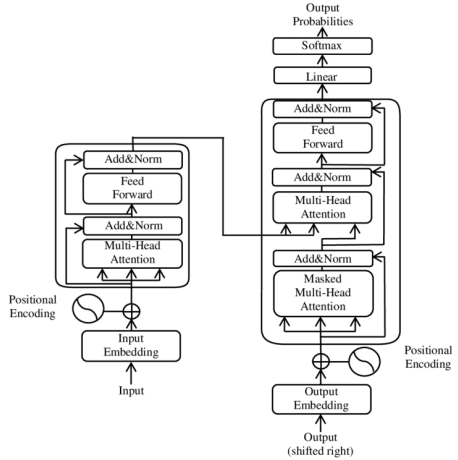
New approach: Sequence-based models

- Use **amino-acid sequence** directly, **without manually deriving physicochemical properties**
- Use **sequence-based deep neural networks**
- Different types of sequence-based models can be investigated, e.g., LSTMs, Bi-LSTM, Transformer



Long-term Short-term Memory (LSTM)

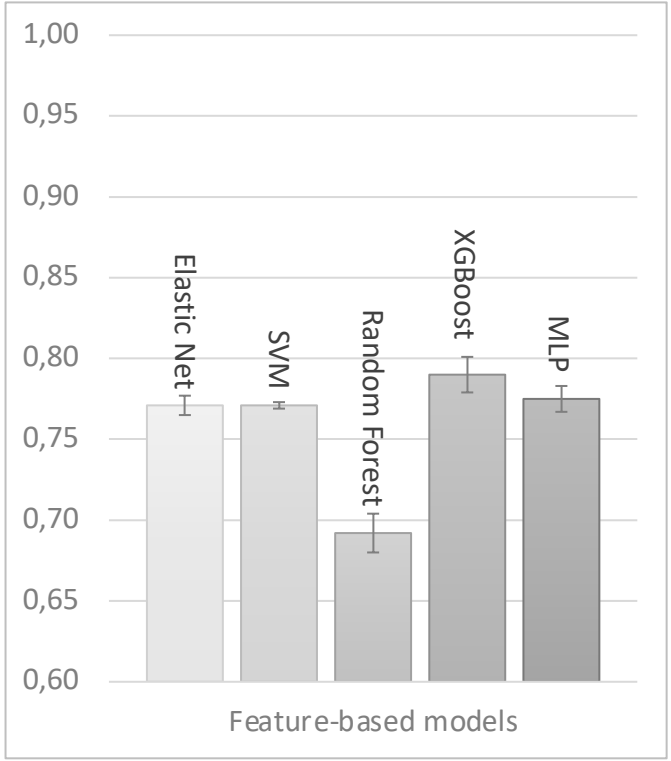


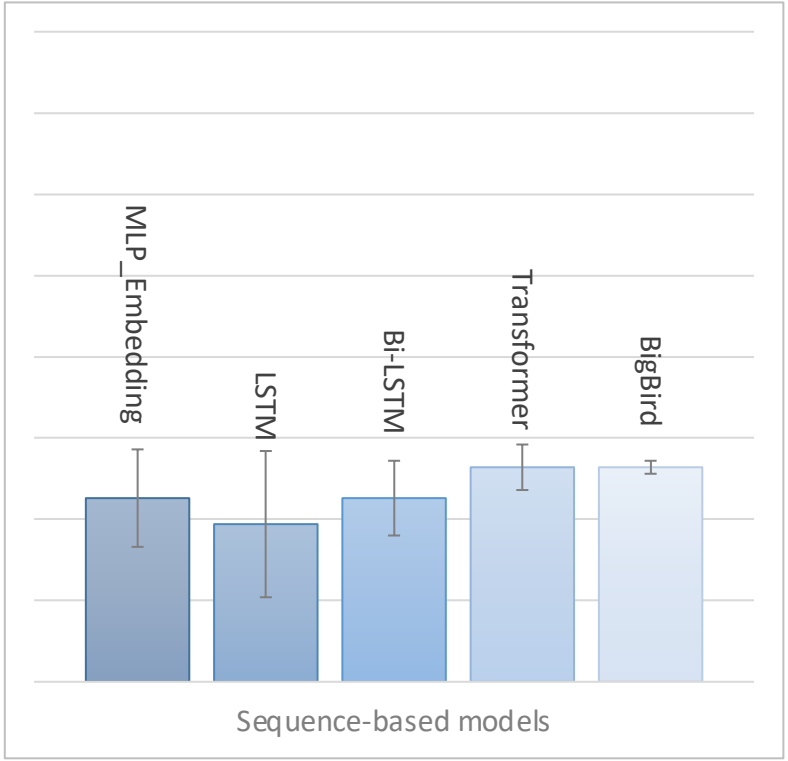


Transformer Model Architecture

Nested cross-validation with **Bayesian hyperparameter optimization**

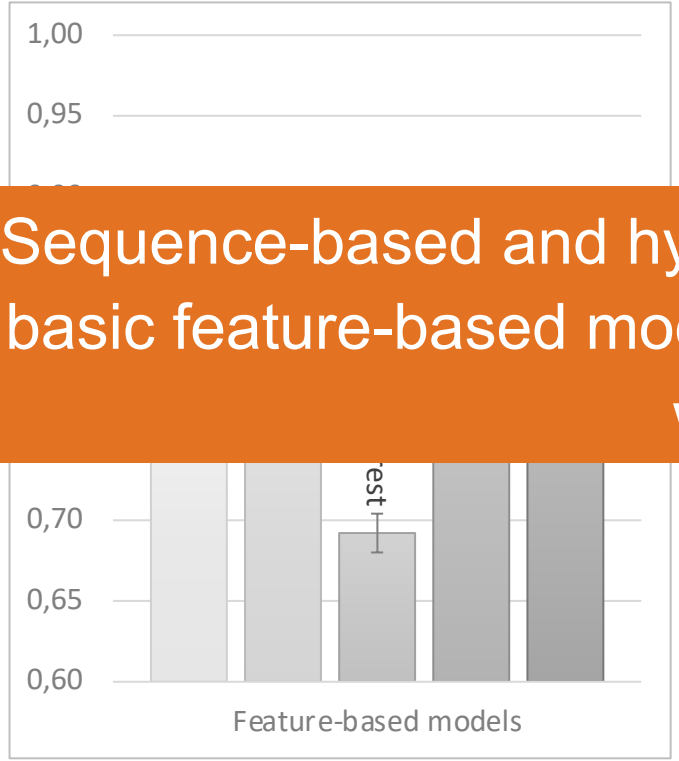
Matthew's Correlation Coefficient (MCC) on test data in nested cross-validation

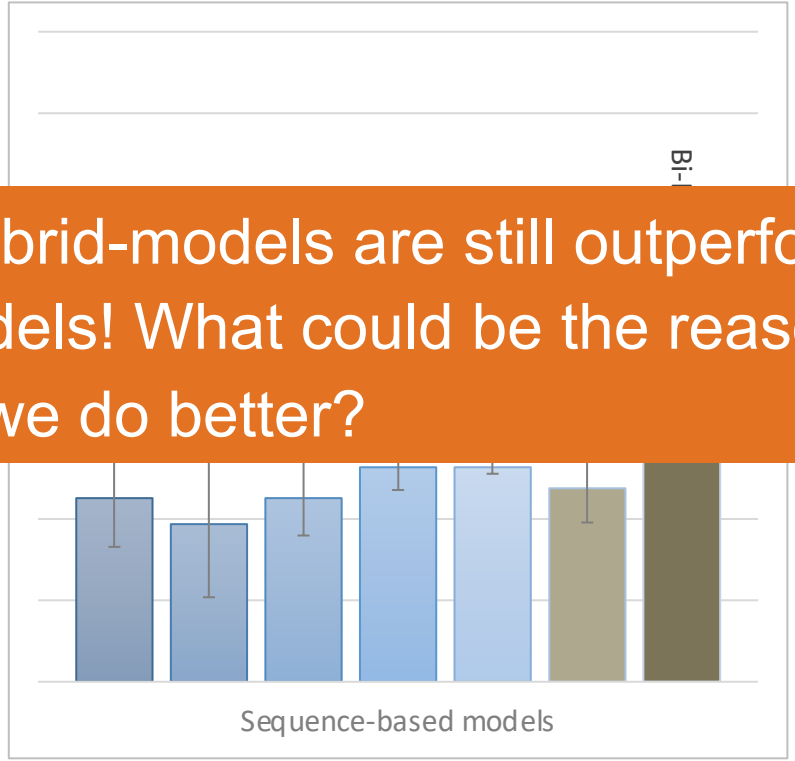




Nested cross-validation with **Bayesian hyperparameter optimization**

Matthew's Correlation Coefficient (MCC) on test data in nested cross-validation





Sequence-based and hybrid-models are still outperformed by basic feature-based models! What could be the reason? Can we do better?

Large Protein Language Models

Original Protein Sequence AESTLGAAAAQSGRYFGTAIASGRLSDSTYTSI...

Masked Protein Sequence AES*LG*AAA***RYF*TAIA*GRLS**TY*SI...

ProtT5XLUniRef50 (Elnaggar et al., 2022)

- Self-supervised training on 50 million UniRef50 protein sequences
- Supercomputer with 5616 GPUs and 1024 TPUs
- ProtT5 outperformed state-of-the-art in terms secondary structures
- Model can learn some of the grammar and language of proteins

Tokenization & Sequence Encoding

Stack of N self-attention transformer layers

Latent amino acid embeddings

Learn to reconstruct sequence AESTLGAAAAQSGRYFGTAIASGRLSDSTYTSI...

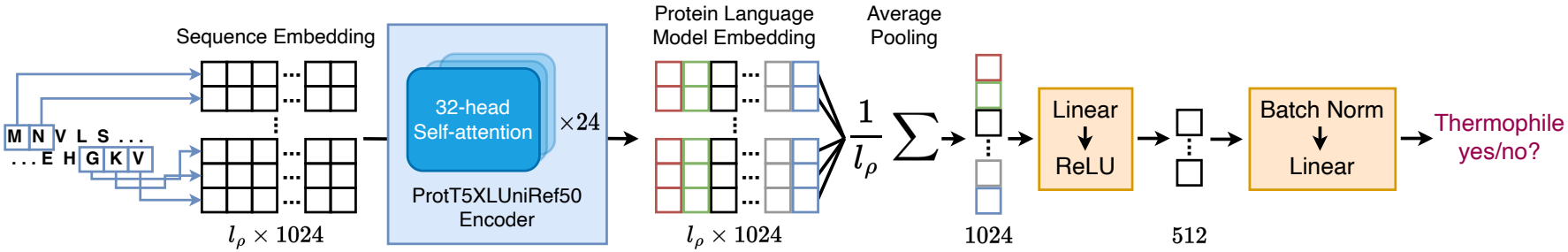
Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., ... & Rost, B. (2021). Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, *44*(10), 7112-7127.

Protein **La**nguage Model-based **Therm**ophilicity Predictor – `ProLaTherm`

- First purely **sequence-based thermophilicity prediction** method
- **`ProLaTherm`** does not rely on **manual feature engineering**
- **`ProLaTherm`** integrates **pretrained embeddings** from **large protein language models** (ProtT5XLUniRef50, Elnaggar et al. 2022)

Florian Haselbeck

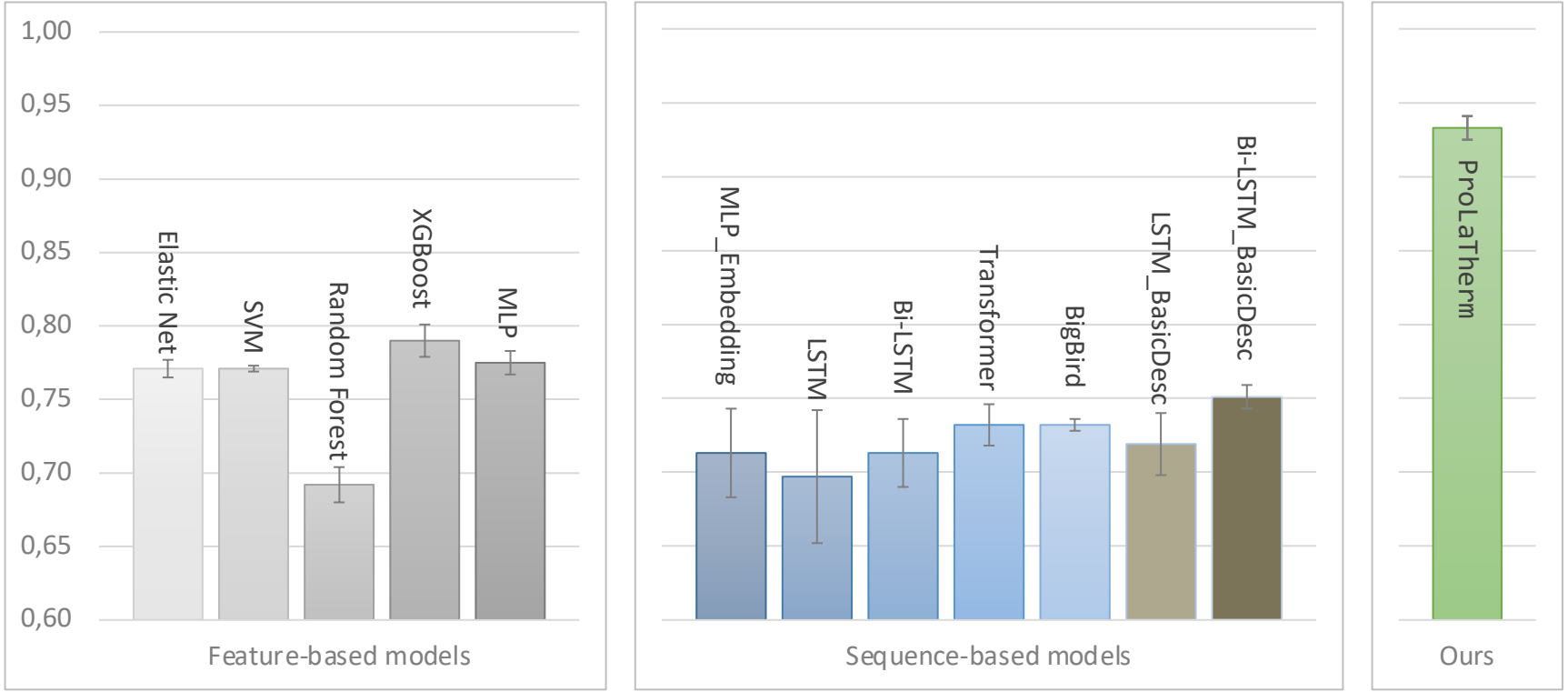
Maura John



Haselbeck F., John M., Zhang Y., Pirnay J., Fuenzalida-Werner J. P., Costa R. D. & Grimm D. G. (2023). Superior Protein Thermophilicity Prediction With Protein Language Model Embeddings, NAR Genomics and Bioinformatics

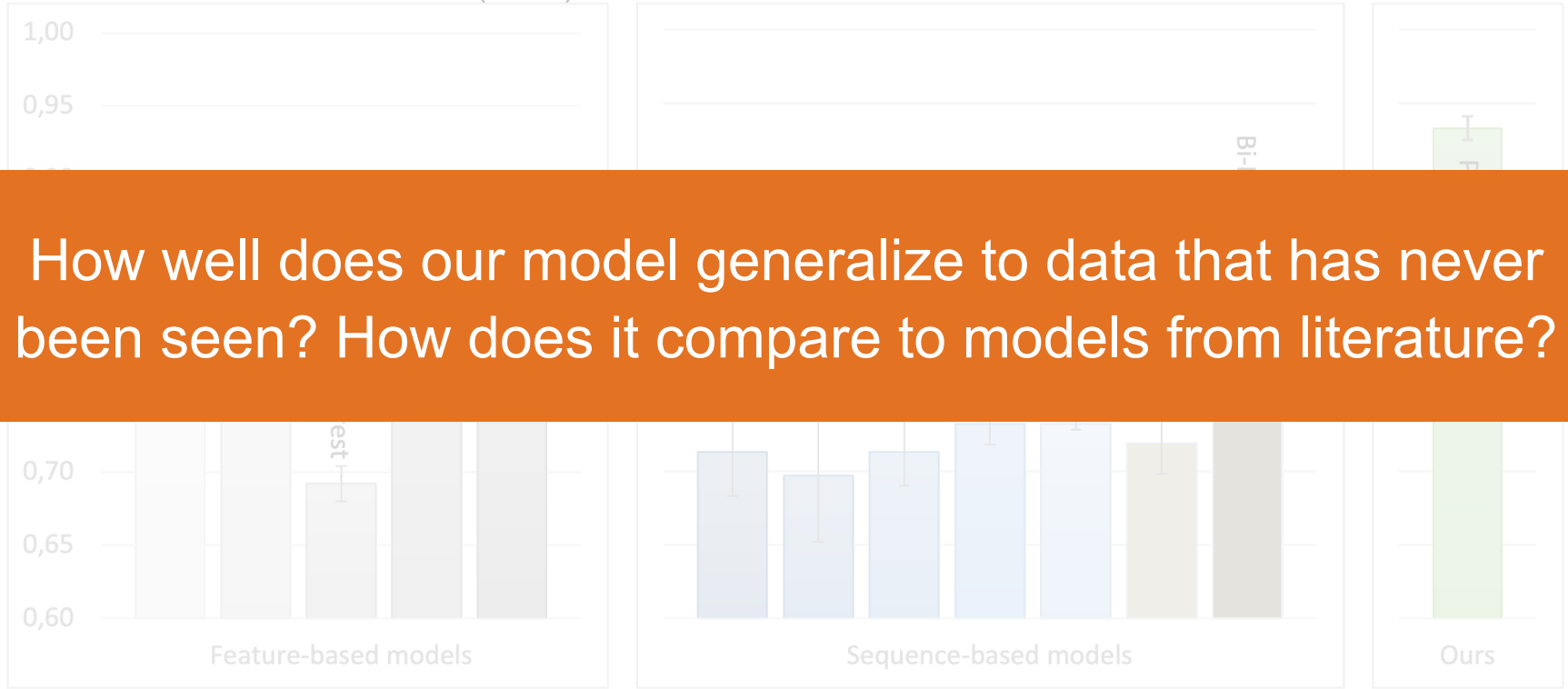
Nested cross-validation with **Bayesian hyperparameter optimization**

Matthew's Correlation Coefficient (MCC) on test data in nested cross-validation



Nested cross-validation with **Bayesian hyperparameter optimization**

Matthew's Correlation Coefficient (MCC) on test data in nested cross-validation



How well does our model generalize to data that has never been seen? How does it compare to models from literature?

Independent Test Data

- We created an **independent test set** to assess the generalization abilities of `ProLaTherm`
- **Not overlapping with data** from tools published in literature
- The **data only contains species and protein sequences** that have not been seen during training (it is **not allowed that different proteins from the same species occur in both, training and testing**)

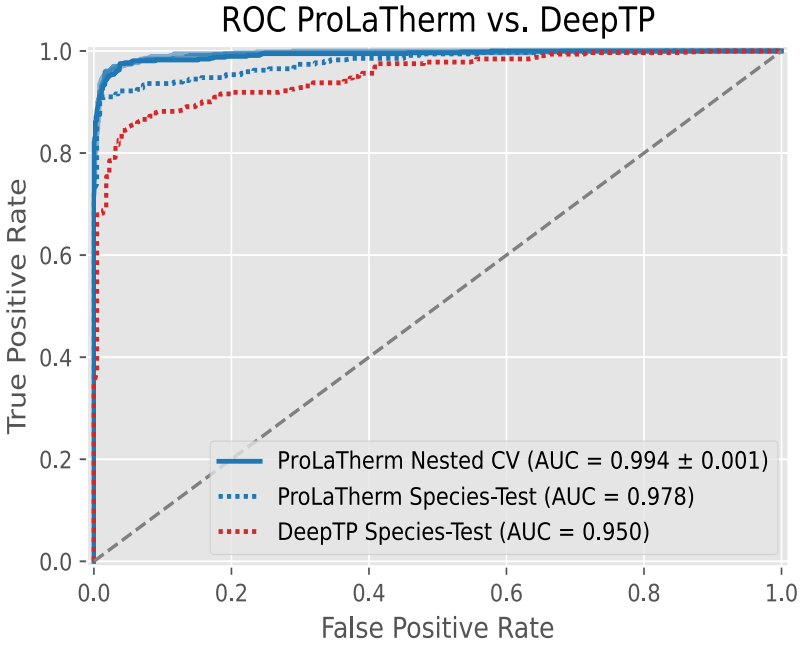
Class	Species	Sequences
Non-thermophilic	75	224
thermophilic	51	345

Species independent test set

Evaluation of ProLaTherm on proteins from species not included in the training

» Independent evaluation of ProLaTherm on novel protein sequences from species not included in the training

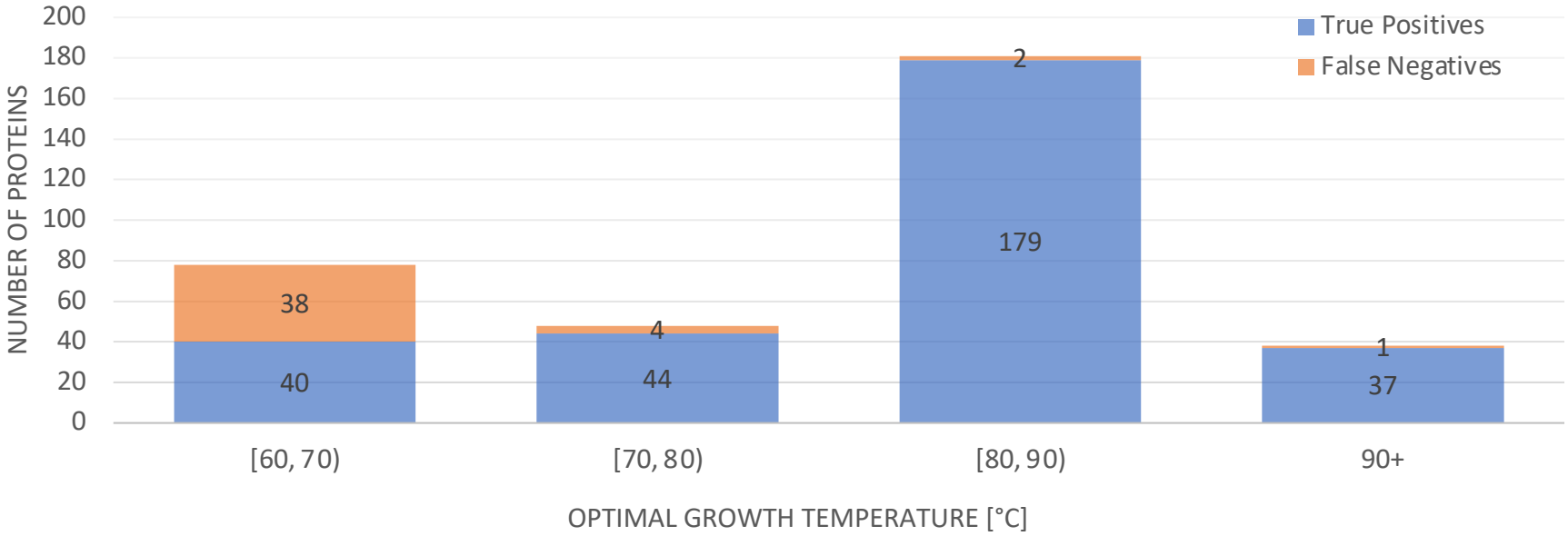
Method	MCC
ThermoPred (Lin and Chen, 2011)	0.635
SCMTPP (Charoenkwan et al. 2021)	0.641
iThermo (Ahmed et al. 2022)	0.637
SAPPHIRE (Charoenkwan et al. 2022)	0.752
DeepTP (Zhao et al. 2023)	0.772
BertThermo (Pei et al. 2023)	0.757
ProLaTherm (ours)	**0.847**



→ **ProLaTherm outperforms the best predictor from the literature by at least 9.3% (DeepTP)**

Prediction Analysis of ProLaTherm

Performance of ProLaTherm on thermophilic species of the independent test set for different optimal growth temperatures



Summary

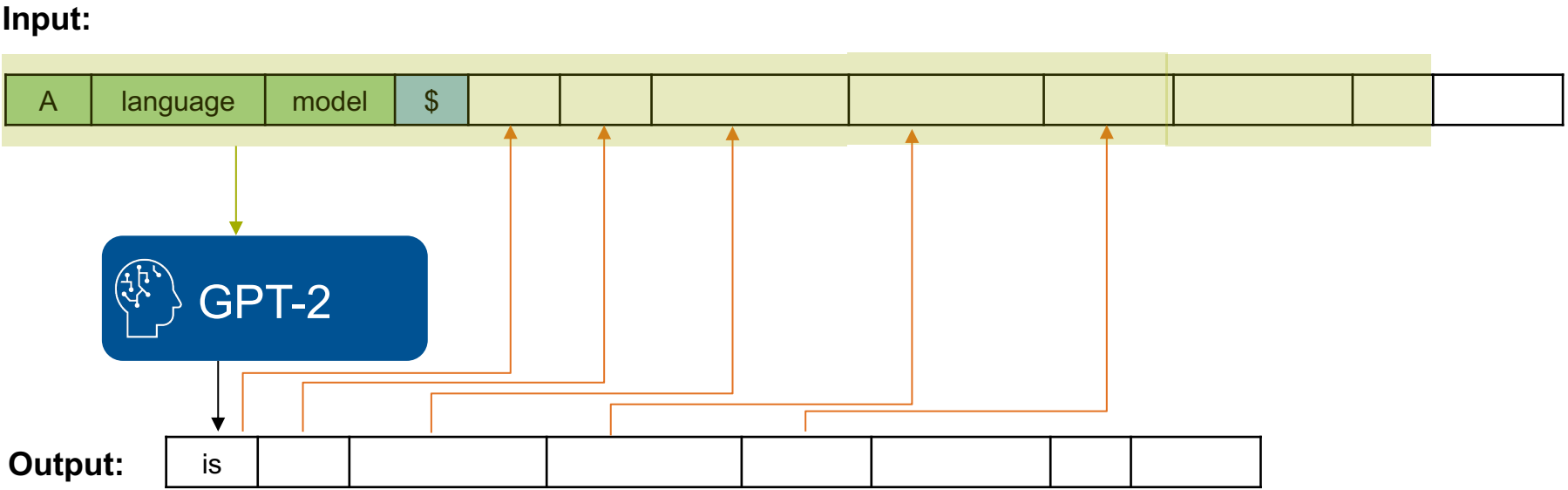
- First purely **sequence-based thermophilicity prediction** method that does not rely on manual feature engineering
- `ProLaTherm` integrates **pre-trained embeddings from protein language models** (ProtT5XLUniRef50, Elnaggar et al. 2022)
- `ProLaTherm` is **superior in thermophilicity prediction** with respect to all comparison partners
- `ProLaTherm` performs very well for proteins with an OGT above 70°C with low false negative rates (below 2.6%)

WEIHENSTEPHAN · TRIESDORF
University of Applied Sciences

TUMCS TUM

Synthetic Protein Design using Generative Machine Learning

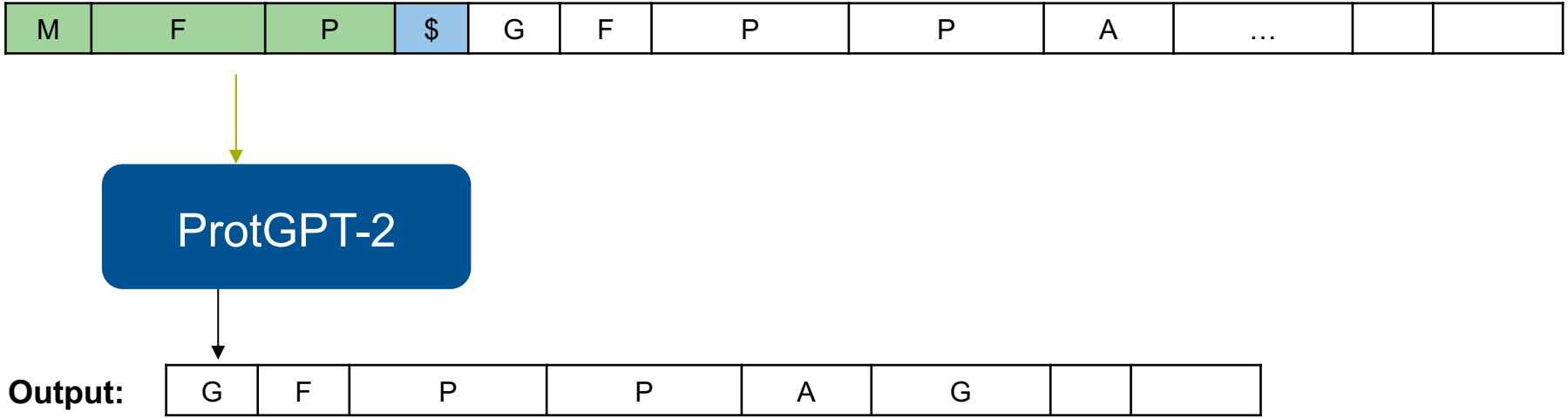
Generative Pretrained Transformer (GPT)



- GPT-2 outputs one token at a time based on a probability
- The generated token is then fed back to the input sequence and is used as new input to the model to generate the next token

Protein Generative Pretrained Transformer (ProtGPT-2)

Input:



- ProtGPT-2 is trained on 50 million protein sequences from Uniref50
- 10% of the sequences were randomly selected as validation set

Ferruz, N., Schmidt, S., & Höcker, B. (2022). ProtGPT2 is a deep unsupervised language model for protein design. *Nature communications*, *13*(1), 4348.

Synthetic Protein Design with `GlycoGPT`

- We used the pretrained **ProtGPT2** and **fine-tuned and retrained** the model using transfer learning on Glycosyltransferase Family 10 (GT10) sequences
- Our adapted model **`GlycoGPT`** is then used to **generate novel** amino-acid **sequences** from the **GT10** family
- We developed **bioinformatics pipeline to evaluate** the generated sequences with respect to plausibility to select promising candidates for **evaluation in the wet-lab** (primary sequence, BLAST similarity, secondary structure, solubility, activity, thermostability and 3D structure using AlphaFold predictions)

Dr. Sara Omranian

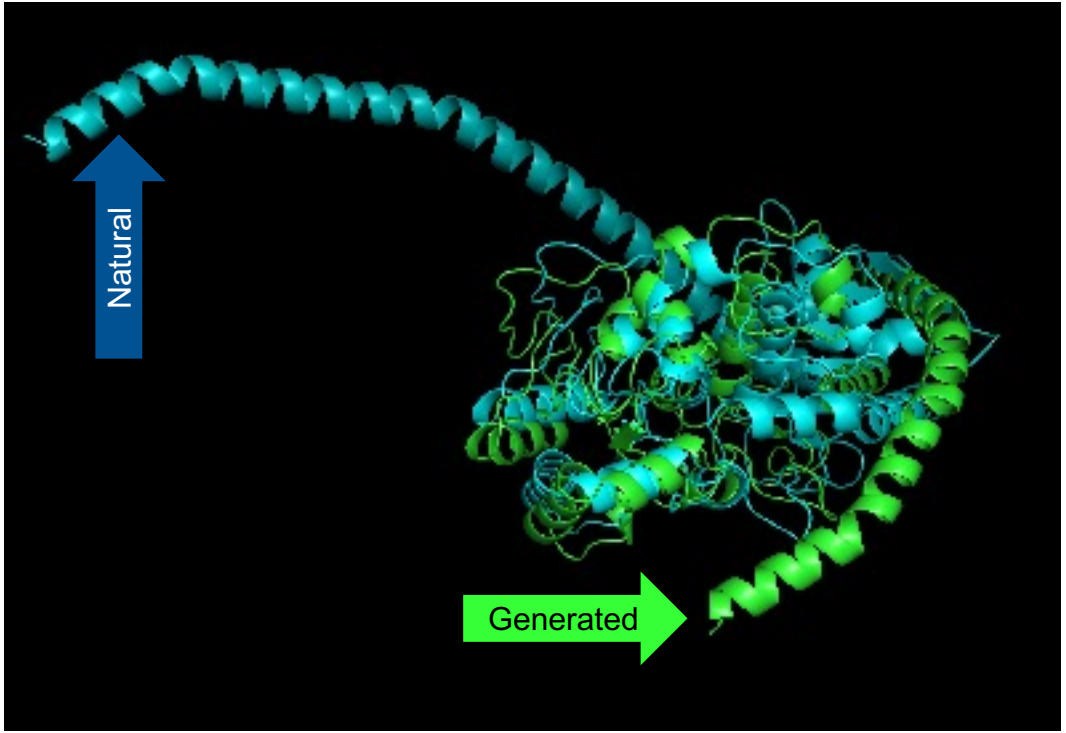
Florian Haselbeck

Sofia Martello

Gefördert durch
Bayerisches Staatsministerium für
Wirtschaft, Landesentwicklung und Energie

Example Protein

GlycoGPT




```
Matrix: EBLOSUM62
Gap penalty: 2.0
Extend penalty: 2.0
Score: 1518.0
Sequence 1 length:328
Sequence 2 length:425
Alignment length: 427
Identity:     284/427 (66.51%)
Similarity:   305/427 (71.43%)
Gaps:         101/427 (23.65%)

1    ------------------------------------------------------------  1
1    MFQPLLDAYTDSTHLDDTTHKPPLNIALANWWPSKNSEKEGFRDFIIHVILKQRYTITLH  60
1    ----------------------------AV--SA-FEVPNFNLFDYAIGFDELDFRDRYL  29
61   QNPNEPSDLVFGNALGQARKILSYQNTKRVFYTGENEAPNFNLFDYAIGFDELDFNDRYL  120
30   RMPLYYAYLHYKALLVNDTTAPYKIKSDTLYTLKKPSHKFKENHPHLCALIHNESDPLKR  89
121  RMPLYYAYLHYKAMLVHDTTAPYKLKPDSLYTLTKPSHKFKENHPMLCALIHNESDPLKR  180
90   GFASFVASNANAPVRNAFYDALNSIEPVTGGGSVKNTLGYNVTNKSEFLSQYKFNLCFEN  149
181  GFVSFVASNPNAPIRNAFYDALNSIEPVTGGGSVKNTLGYKVKNKMEFLSQYKFNLCFEN  240
150  SQGYGYVTEKNI-ESMMAGSIPVYWGSPSVAKDFNPKSFVNVHDFKNFDEAIDYVRYLHT  208
241  SQGYGYVTEKCILDAYFSHTIPIYWGSPSVAKDFNPKSFVNVHDFKNFDEAIDHIRYLHA  299
209  HPNAYLEMLYENPLNEIDGKAGFYQNLSFKKILDFFKTILENDTIYHNNPFIFYRDLNEP  268
300  HQNAYLDMLYENPLNTLNGKAGFHQDLSFQKILDFFKTILENDTIYH-NPSALYRDLNEP  358
269  LVSVDDLRV-------NYDDLRRDHERLLSKATPLLELSQNTSFKIYRKAYQKSLPLLRA  321
359  LVSVDDLRVNYDDLRINYDDLRRDHERLLSKATPLLELSQNTSFKIYRKAYQKSLPLLRT  418
322  IRRWVKK                                                       328
419  IRRWVKK                                                       425
```

Synthetic Protein Design with GlycoGPT

- We have started to develop `GlycoGPT`, a generative machine learning model for synthetic protein design of GT10 sequences

- *In-silico* evaluation of generated sequences is rather difficult → the next step is to **evaluate the generated sequences in the laboratory**

- Adding constraints to the model architectures to allow the generation of proteins with **specific functions**

WEIHENSTEPHAN · TRIESDORF
University of Applied Sciences

SynBiofoundry @TUM

TUMCS

TUM

Acknowledgements

Contact Information

Prof. Dr. Dominik Grimm

dominik.grimm@tum.de

http://bit.cs.tum.de/

@dg_grimm

GrimmLab Team

Josef Eiglsperger
Nikita Genze
Florian Haselbeck
Maura John
Sofia Martello
Jonathan Pirnay
Krystian Budkiewicz
Maximilian Wirth
Anna Fischer

Collaborations for these Projects

Volker Sieber
Ruben Costa

Thanks for your attention!

GrimmLab **Team**

GrimmLab **Funding**

Funded by
DFG Deutsche Forschungsgemeinschaft
German Research Foundation

European Innovation Council

Bayerisches Staatsministerium für Ernährung, Landwirtschaft und Forsten

Gefördert durch
Bayerisches Staatsministerium für Wirtschaft, Landesentwicklung und Energie

Gefördert durch
Bundesministerium für Ernährung und Landwirtschaft

Projektträger
Bundesanstalt für Landwirtschaft und Ernährung

aufgrund eines Beschlusses des Deutschen Bundestages

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung