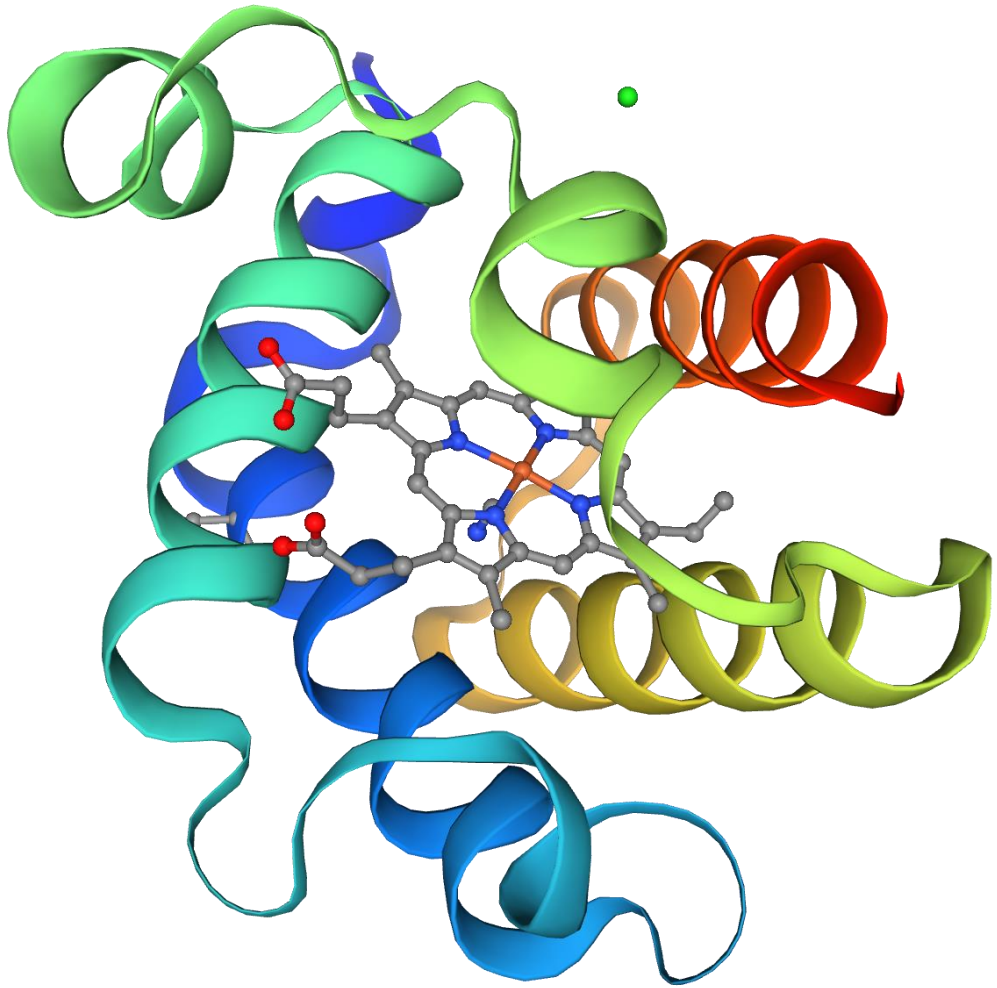
University of Basel

Hyper-Parameter Optimization for Remote Homology Detection with Protein Language Models

Master Thesis, Laura Maria Engist <l.engist@unibas.ch>

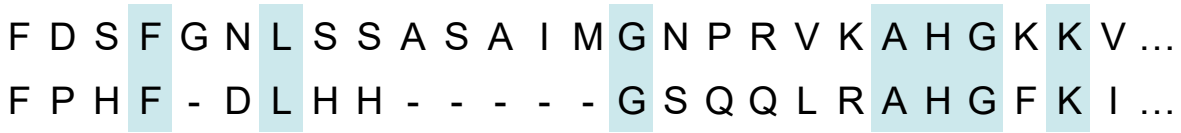
Supervisors and Examiners: Prof. Dr. Malte Helmert, Dr. Janani Durairaj, Dr. Florian Pommerening

Final Presentation, September 5, 2025



NAPYEAIGEELLSQLVDTFYERVASHPLLKPIFPSDLTETARKQKQFLTQYLGGPPLYTEEHGHPMLRARHLPFPITNERADAWLSCMKDAMDHVGLEGEIREFLFGRLELTARHMVNQ

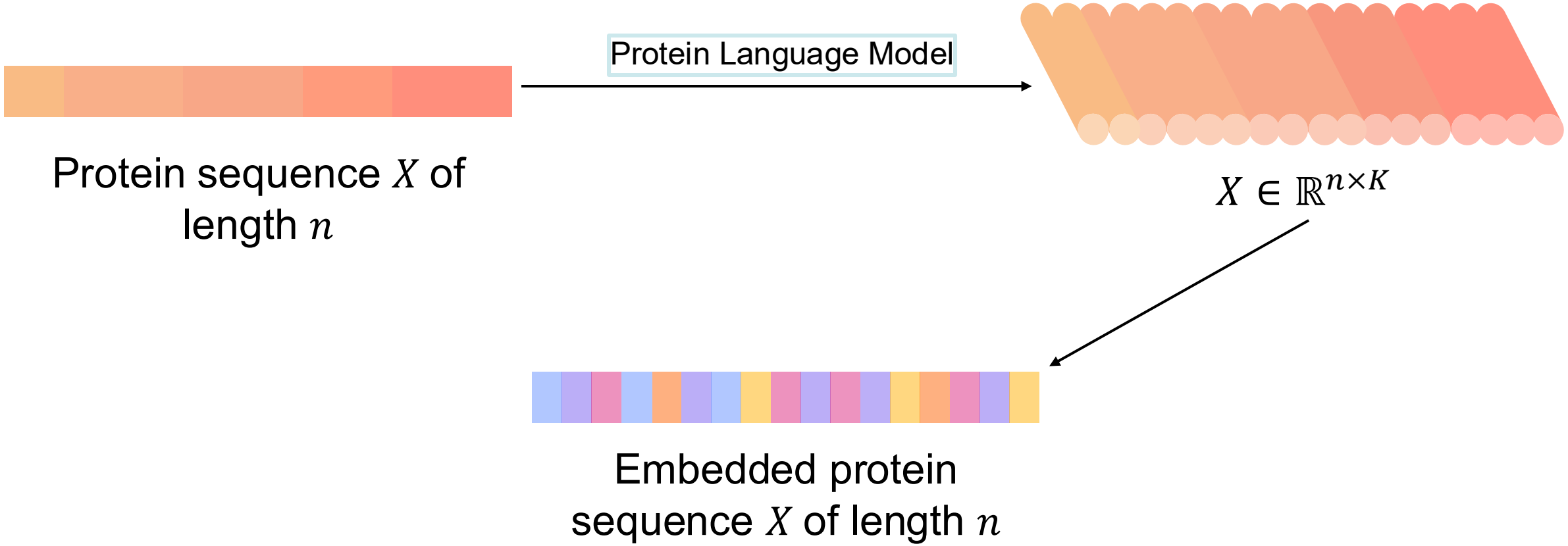
Source: https://swissmodel.expasy.org/templates/1ux8



- **Protein Sequence Alignment**
  - Process to align sequences to highlight similarities

- **Protein Sequence Homology Search**
  - Search for sequence similarity
  - Uncover evolutionary relationships between proteins

Source: https://www.researchgate.net/figure/A-multiple-sequence-alignment-of-protein-sequences-Gaps-are-introduced-to-achieve_fig4_254860534


Protein Language Model
Protein sequence X of length n
$X \in \mathbb{R}^{n \times K}$
Embedded protein sequence X of length n

Main Questions

1) Are sequence alignments significantly affected by the choice of gap penalties?

F D S F G N L S S A S A I M G N P R V K A H G K K V ...
F P H F - D L H H - - - - - G S Q Q L R A H G F K I ...

2) What parameter values work well for configuring the procedure to compute discrete embedded sequences?

Agenda

1	Hyper-Parameter Optimization: SMAC3
2	Optimization Pipelines and Scoring Metrics
3	Experiments and Results
4	Conclusions and Outlook

Hyper-Parameter Optimization: SMAC3

Hyper-Parameter Optimization: SMAC3

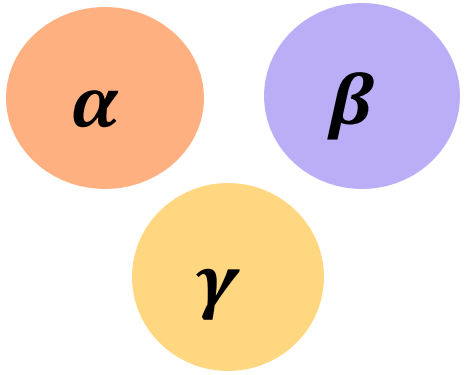
- Versatile package for hyper-parameter optimization (Python3 and C++)
- Developed by the AutoML Groups of the Universities of Hannover and Freiburg
- **Bayesian optimization** idea

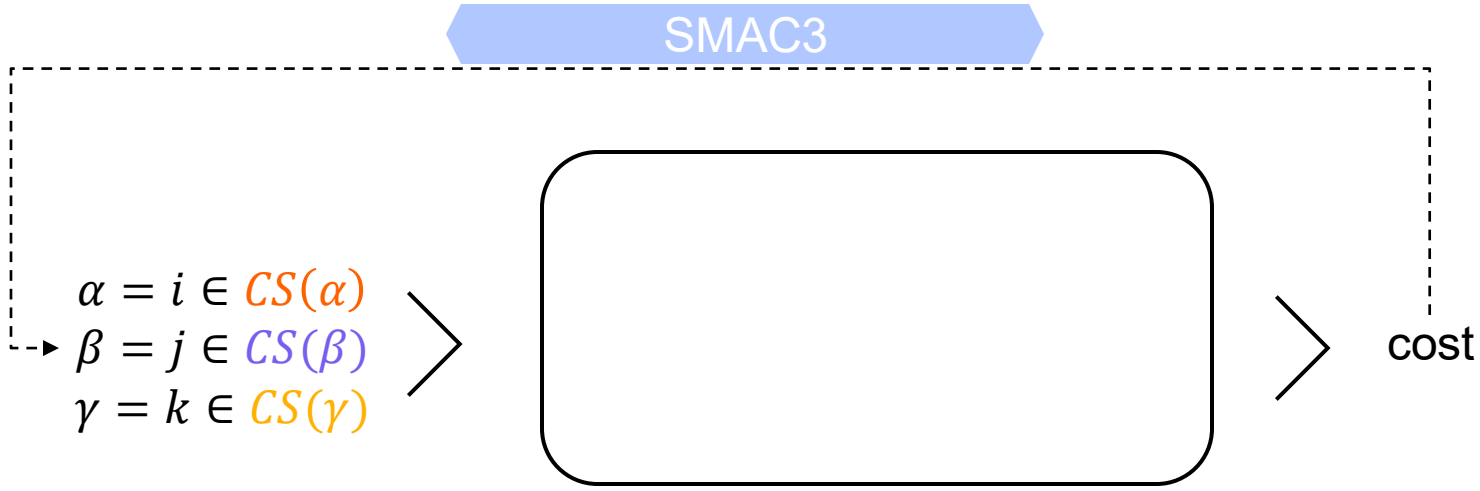
  → Exploration vs. exploitation
- **Random forests**

  → Decision trees
- Logs

Sources:
https://github.com/automl/SMAC3
M. Lindauer, K. Eggensperger, M. Feurer, A. Biedenkapp, and D. Deng, “SMAC3: A Versatile Bayesian Optimization Package for Hyperparameter Optimization,” Journal of Machine Learning Research, vol. 23, pp. 1–9, 2022.

Hyper-Parameter Optimization: SMAC3





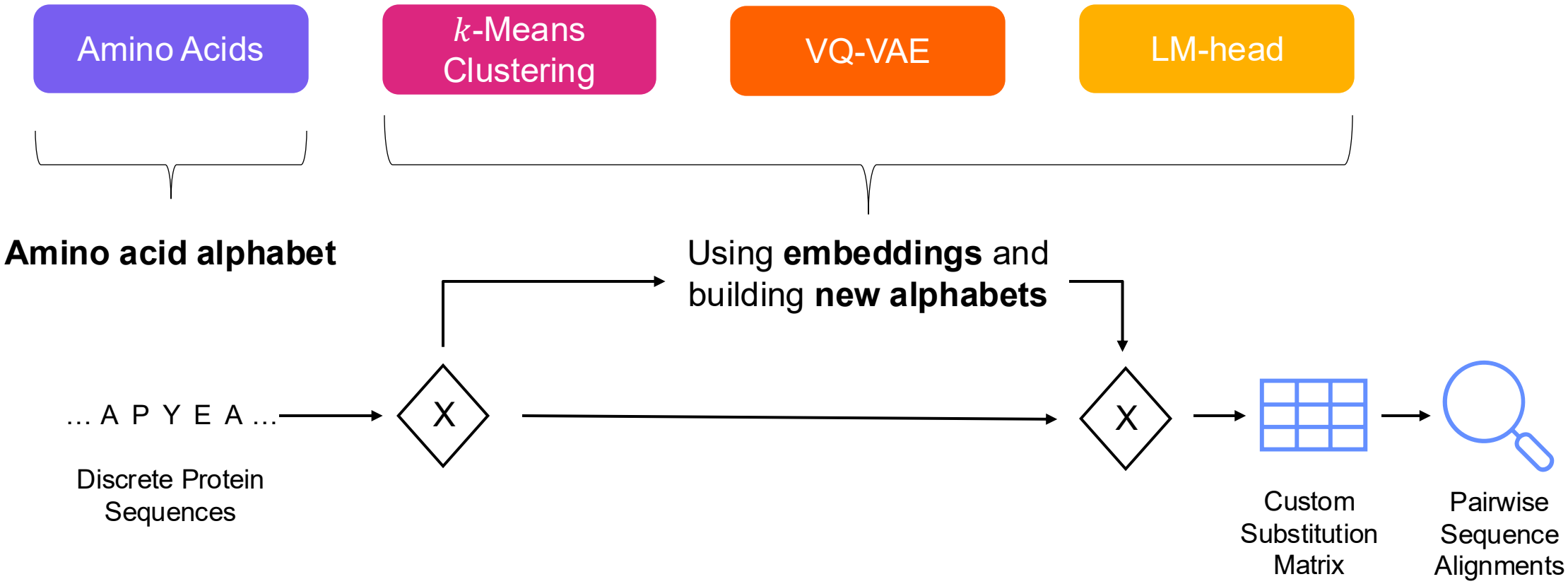
Parameters α, β, and γ, and their configuration spaces (CS)

Agenda

1	Hyper-Parameter Optimization: SMAC3
2	Optimization Pipelines and Scoring Metrics
3	Experiments and Results
4	Conclusions and Outlook

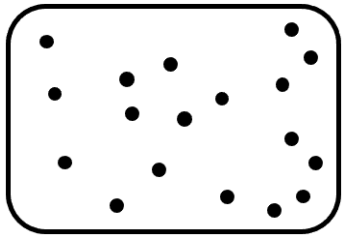
Optimization Pipelines

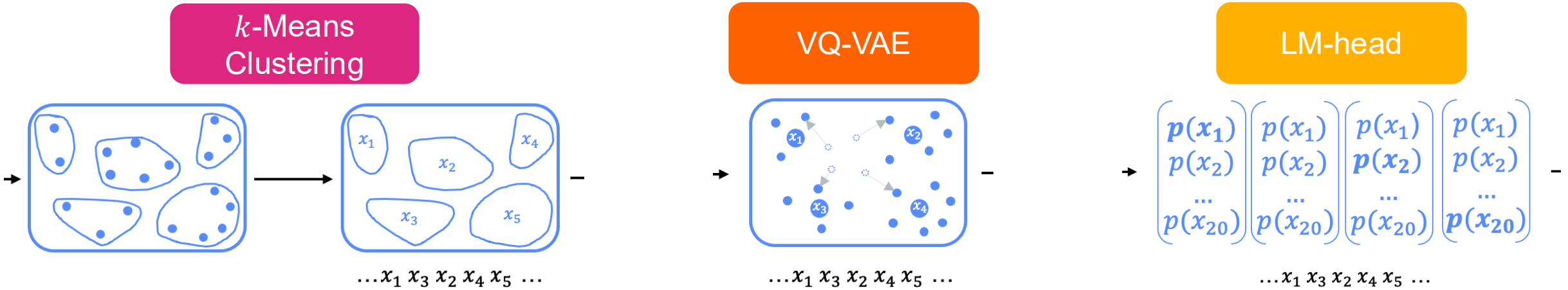
Optimization Pipelines



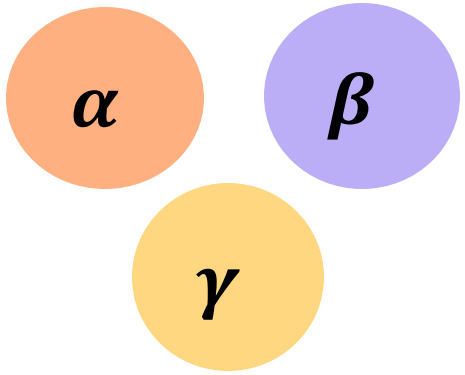
Optimization Pipelines

- Start with pre-computed Protein Language Model embeddings
- Difference: discretization approach

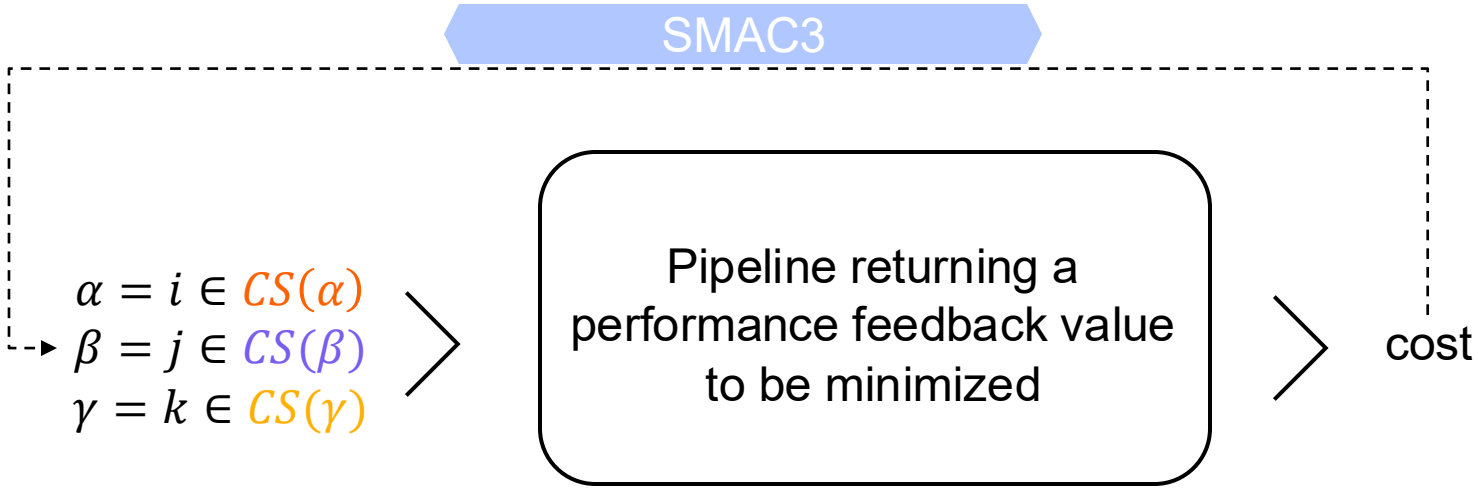




Hyper-Parameter Optimization: SMAC3



Parameters α, β, and γ, and their configuration spaces (CS)

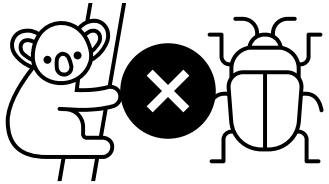


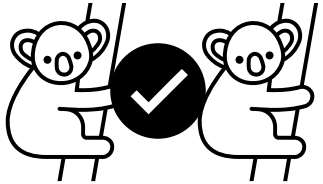
Scoring Metrics

Scoring Metrics

- **Alignment** Quality

- **Identification** Quality





Agenda

1	Hyper-Parameter Optimization: SMAC3
2	Optimization Pipelines and Scoring Metrics
3	Experiments and Results
4	Conclusions and Outlook

Experiments and Results

Experimental Setup

Training and **validation data**: different subsets of the protein sequence databases SCOPe and Pfam

- **Training** data: **24’627** pairs of sequences
- **Validation** data: **11’549** pairs of sequences
- Sequences of length up to 1’024 characters

Two experiments

- Experiment 1: Influence of gap penalties: gap-open (go) and gap-extension (ge)
- Experiment 2: Hyper-parameter optimization

Experimental Setup

Average duration of **one run**:

- Amino Acid: 3 minutes
- k-Means: 7 minutes
- VQ-VAE: 7 hours and 52 minutes
- LM-head: 2 hours and 48 minutes

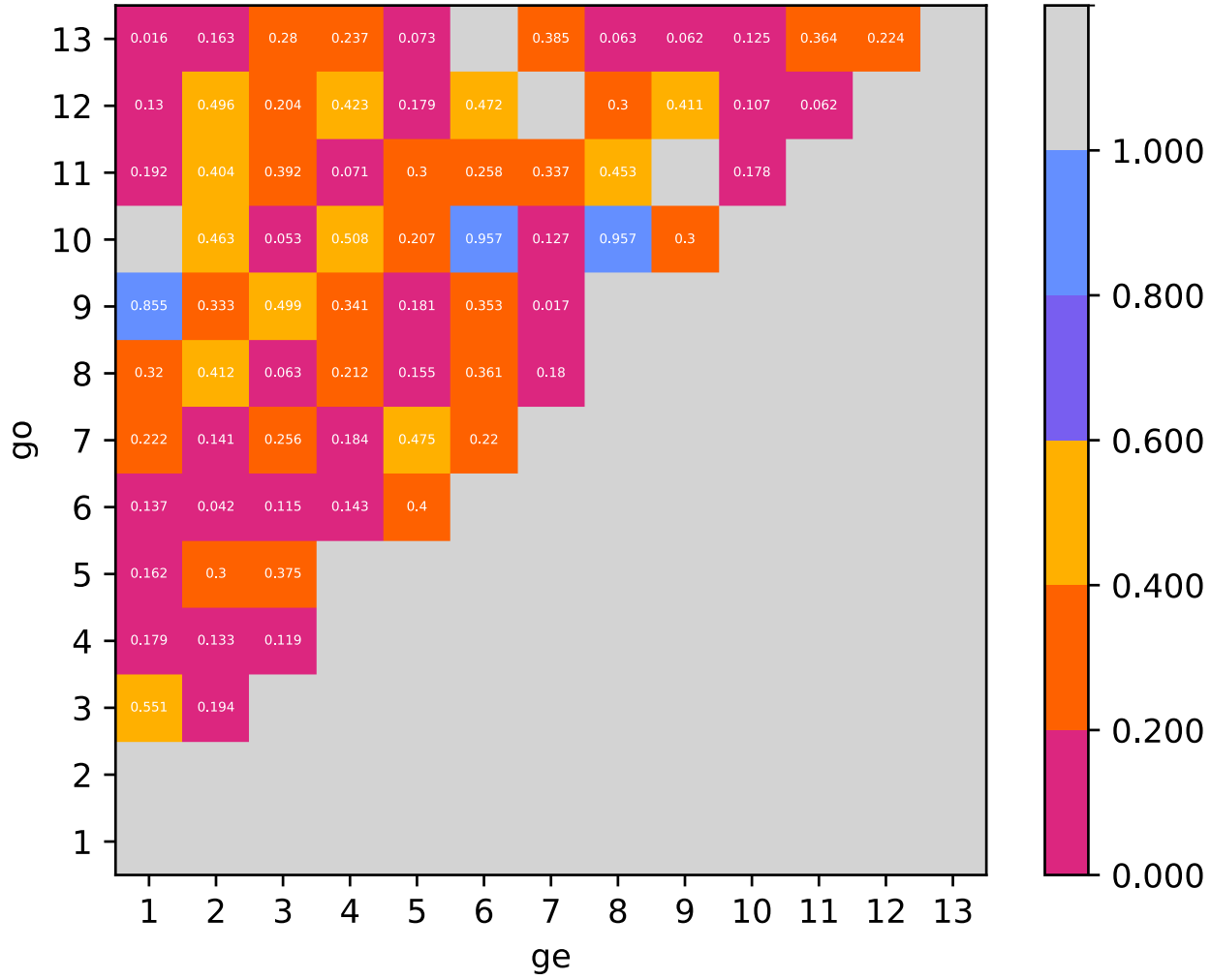
→ Of course, we parallelized the runs!

Main Questions

1) **Are sequence alignments significantly affected by the choice of gap penalties?**

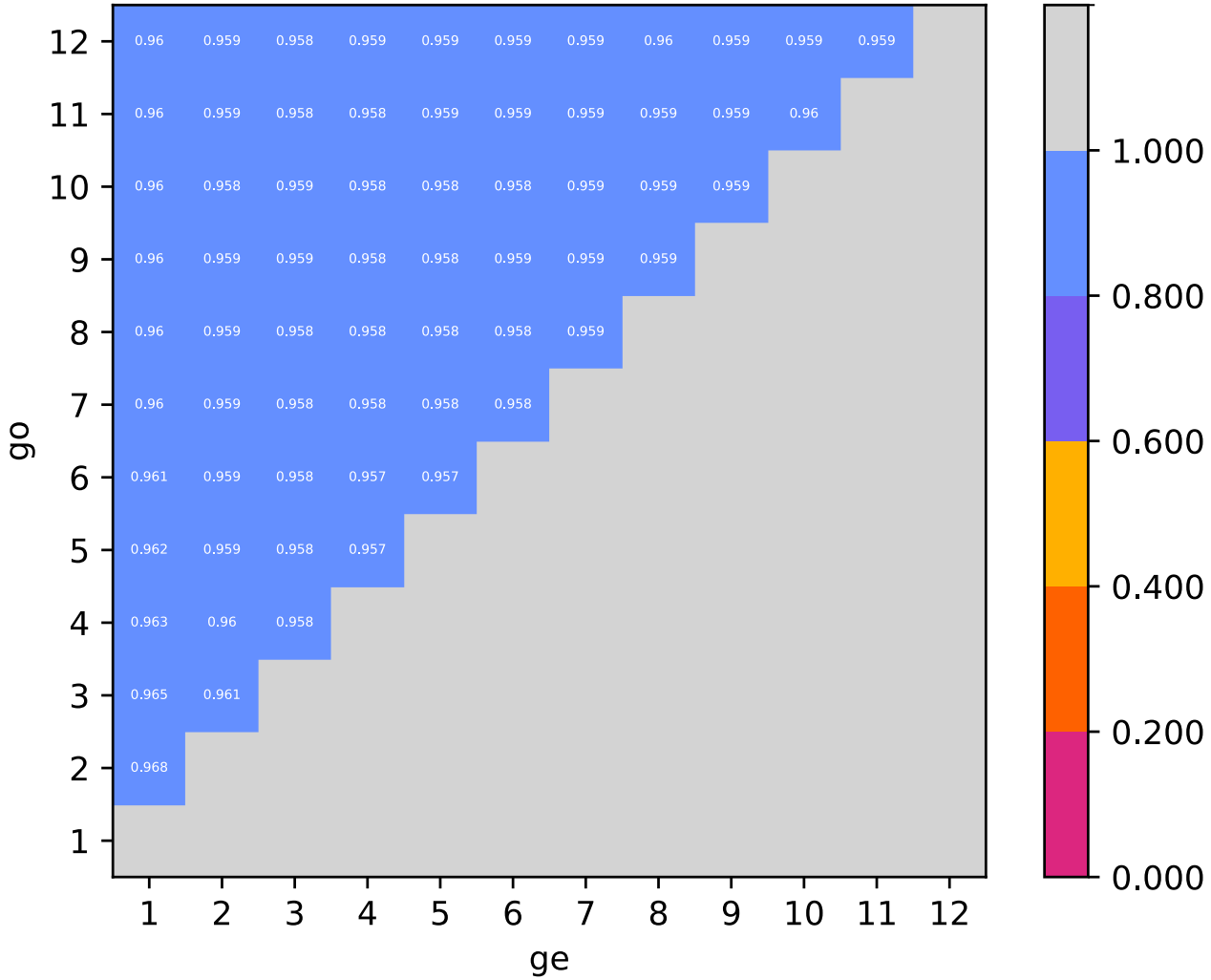
2) What parameter values work well for configuring the procedure to compute discrete embedded sequences?

Experiment 1:
Influence of Gap Penalties – Alignment Quality



Largely differing cost values
→ **significant influence**

Experiment 1:
Influence of Gap Penalties – Identification Quality

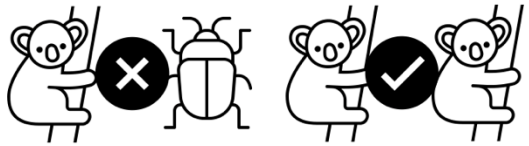


Values vary in a maximum range of 0.11
→ **no significant influence**

Main Questions

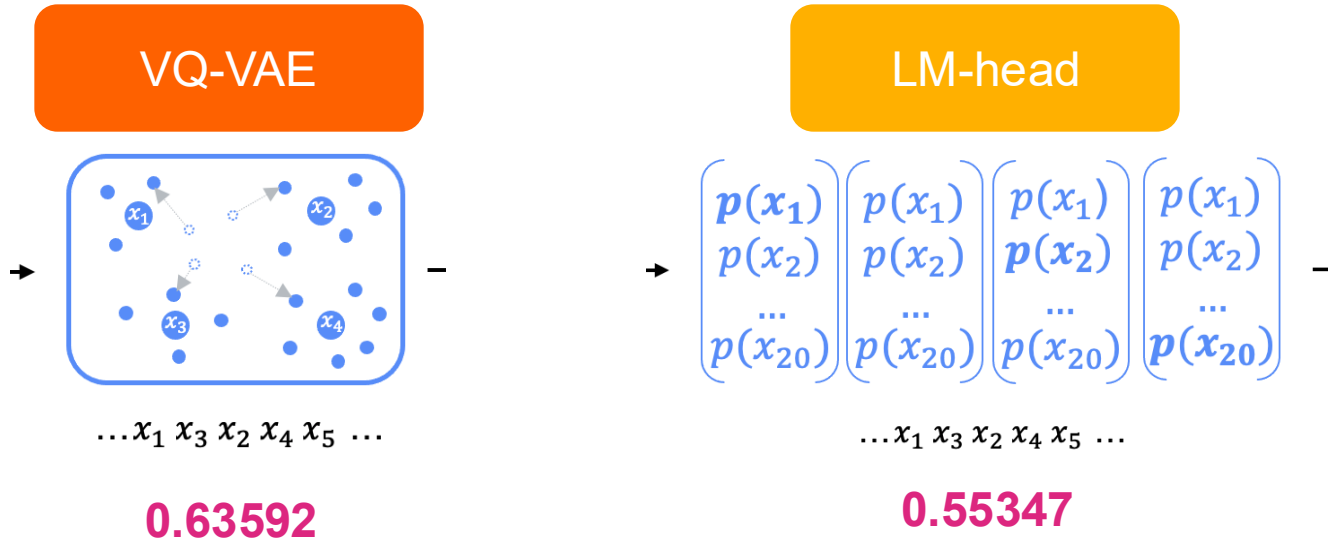
1) Are sequence alignments significantly affected by the choice of gap penalties?
   - **Significant influence** on the quality of **alignments**
   - **No significant effect** on the **identification** of evolutionary background

2) **What parameter values work well for configuring the procedure to compute discrete embedded sequences?**



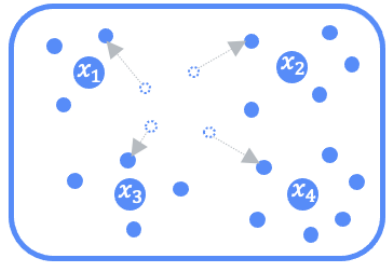
Baseline – Reference Cost

- Reference values to evaluate how much we can improve the identification quality with our optimization



VQ-VAE

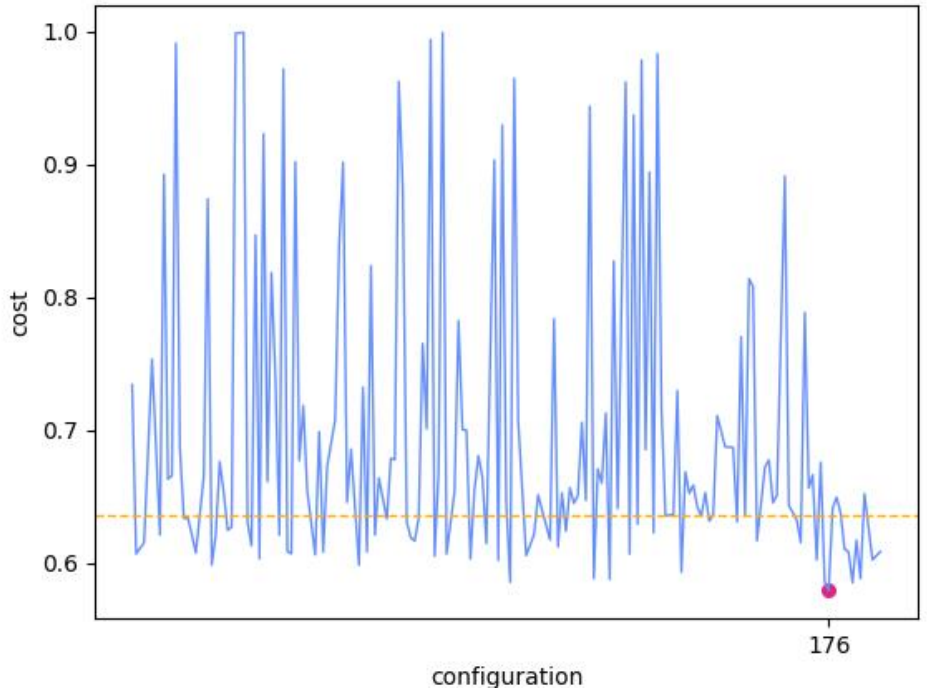
Experiment 2: VQ-VAE Pipeline

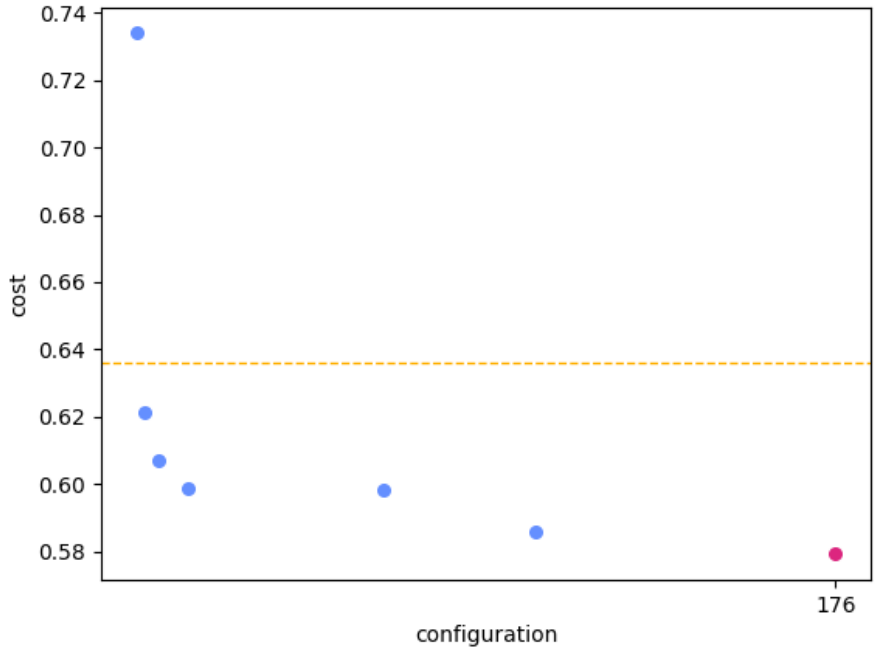


Baseline Cost 0.63592

Setup Start pipeline for 1000 runs (configurations sampled). Stop at time limit.

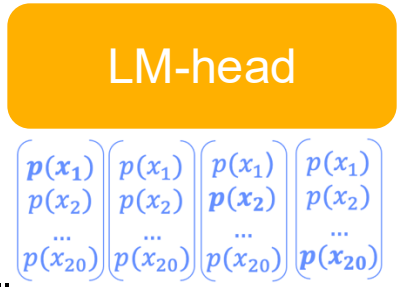
Result Best found configuration number 176 with a cost of **0.57011**





TOD
O
Layoi
ut

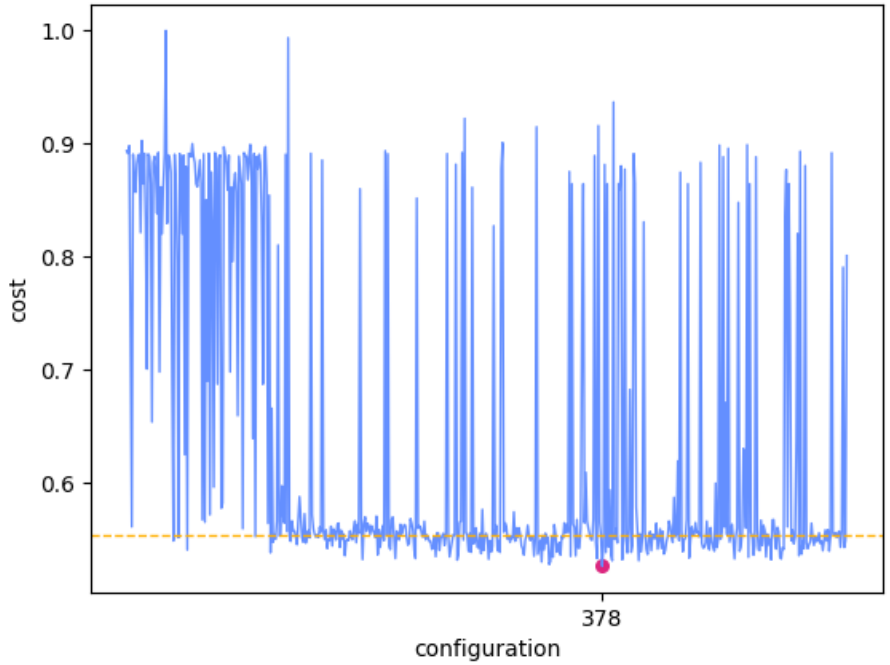
Experiment 2: LM-head Pipeline

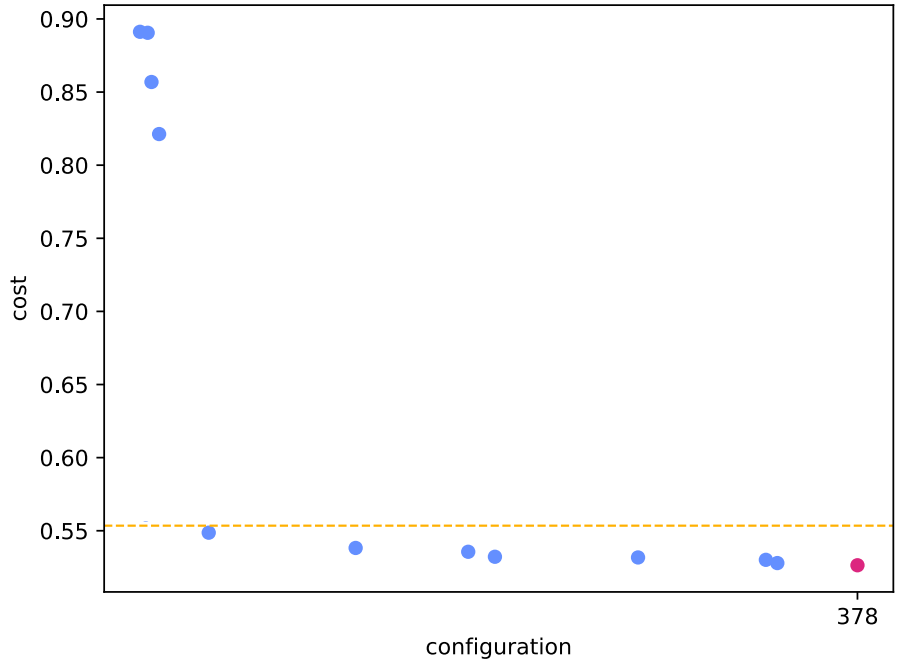


Baseline Cost 0.55347

Setup Start pipeline for 1000 runs (configurations sampled). Stop at time limit.

Result Best found configuration number 378 with a cost of **0.5264**





Main Questions

1) Are sequence alignments significantly affected by the choice of gap penalties?
   - **Significant influence** on the quality of **alignments**
   - **No significant effect** on the **identification** of evolutionary background

2) What parameter values work well for configuring the procedure to compute discrete embedded sequences?
   - Identified configurations for both neural network models
   - **Improved** the **identification** of evolutionary background up to ~ **7%**

Agenda

1	Hyper-Parameter Optimization: SMAC3
2	Optimization Pipelines and Scoring Metrics
3	Experiments and Results
4	Conclusions and Outlook

Conclusions and Outlook

Conclusions

1) Are sequence alignments significantly affected by the choice of gap penalties?
    - **Significant influence** on the quality of **alignments**
    - **No significant effect** on the **identification** of evolutionary background

2) What parameter values work well for configuring the procedure to compute discrete embedded sequences?
    - Identified configurations for both neural network models
    - **Improved** the **identification** of evolutionary background up to ~ **7%**

Conclusions

Limitations:

- Optimization process relies on performance and availability of the compute cluster

  → Influence on experimental throughput

- Fixed databases

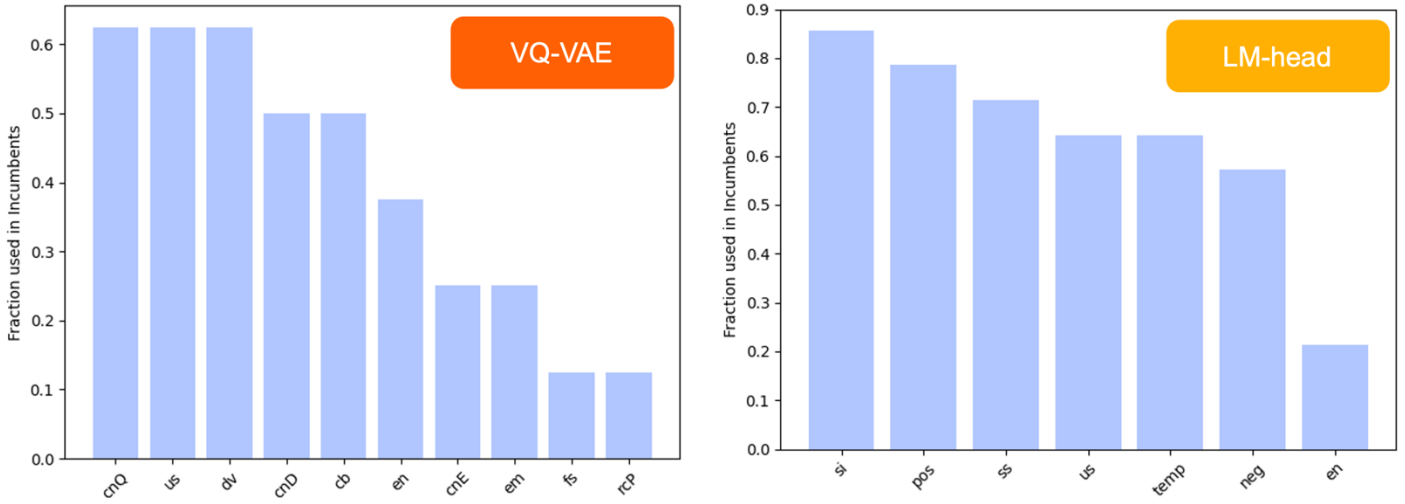
  → Potential limit of generalizability of our results

- Focused on one hyper-parameter optimization framework

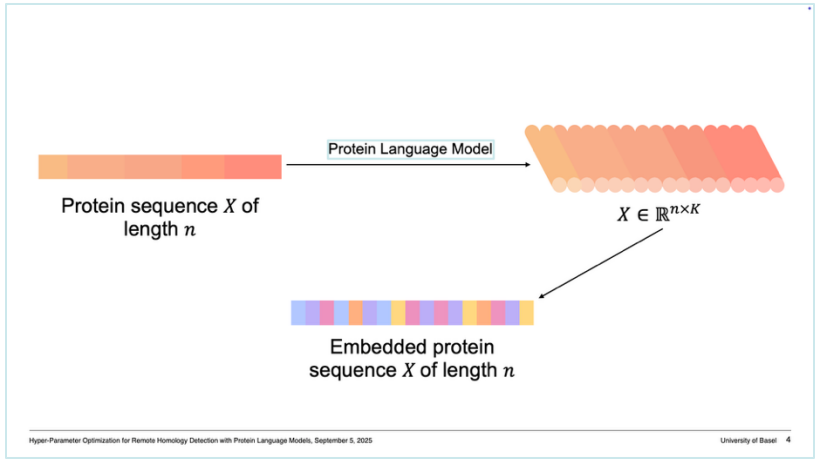
  → Potential improvement with alternative frameworks

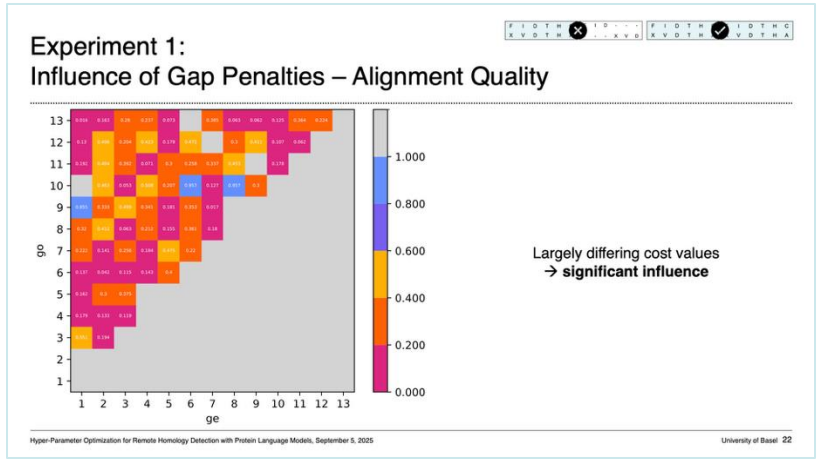
Outlook

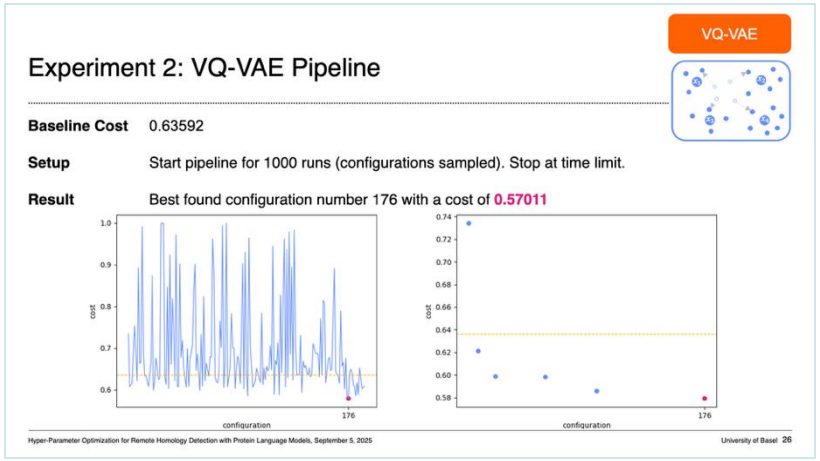
- Best configurations
  - → Translate entire databases of protein sequences
  - → Enable more effective large-scale search
- Parameter importance
  - → Restrict the optimization process to a smaller subset of hyper-parameters
  - → Potentially decrease number of runs to find improvements
- Alternative optimization frameworks

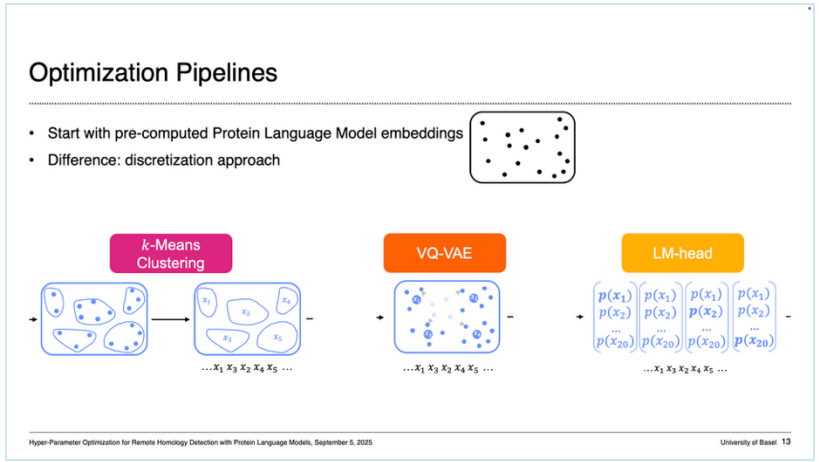


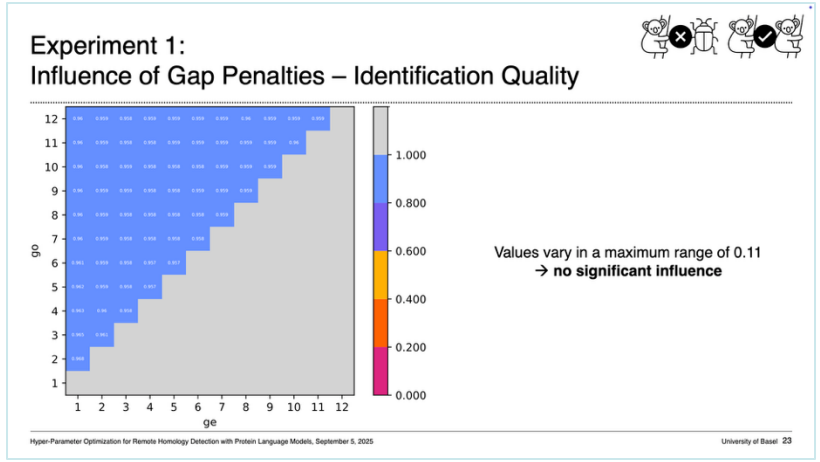
Key Points

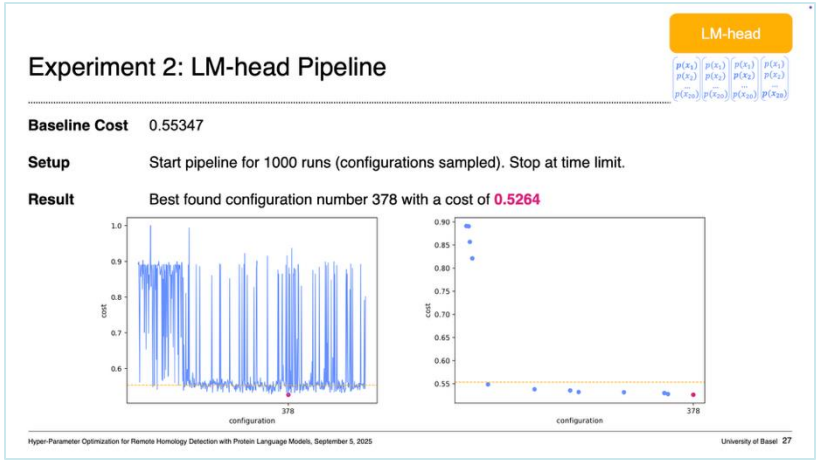

Protein Language Model
Protein sequence X of length n
$X \in \mathbb{R}^{n \times K}$
Embedded protein sequence X of length n
Hyper-Parameter Optimization for Remote Homology Detection with Protein Language Models, September 5, 2025
University of Basel 4


Experiment 1:
Influence of Gap Penalties – Alignment Quality
go
ge
Largely differing cost values
→ significant influence
Hyper-Parameter Optimization for Remote Homology Detection with Protein Language Models, September 5, 2025
University of Basel 22


VQ-VAE
Experiment 2: VQ-VAE Pipeline
Baseline Cost 0.63592
Setup Start pipeline for 1000 runs (configurations sampled). Stop at time limit.
Result Best found configuration number 176 with a cost of 0.57011
cost
configuration
Hyper-Parameter Optimization for Remote Homology Detection with Protein Language Models, September 5, 2025
University of Basel 26


Optimization Pipelines
Start with pre-computed Protein Language Model embeddings
Difference: discretization approach
k-Means Clustering
VQ-VAE
LM-head
Hyper-Parameter Optimization for Remote Homology Detection with Protein Language Models, September 5, 2025
University of Basel 13


Experiment 1:
Influence of Gap Penalties – Identification Quality
go
ge
Values vary in a maximum range of 0.11
→ no significant influence
Hyper-Parameter Optimization for Remote Homology Detection with Protein Language Models, September 5, 2025
University of Basel 23


LM-head
Experiment 2: LM-head Pipeline
Baseline Cost 0.55347
Setup Start pipeline for 1000 runs (configurations sampled). Stop at time limit.
Result Best found configuration number 378 with a cost of 0.5264
cost
configuration
Hyper-Parameter Optimization for Remote Homology Detection with Protein Language Models, September 5, 2025
University of Basel 27