RTH
CENTER FOR NON-CODING RNA IN TECHNOLOGY AND HEALTH

ENSAFE

UNIVERSITY OF COPENHAGEN

Peptide pesticides

off-target binding prediction

Vi Tran, PhD student

Center for non-coding RNA in Technology and Health (RTH)

Faculty for Health and Medical Sciences

University of Copenhagen

Why does off-target assessment of biopesticides matter?

Why does off-target assessment of biopesticides matter?

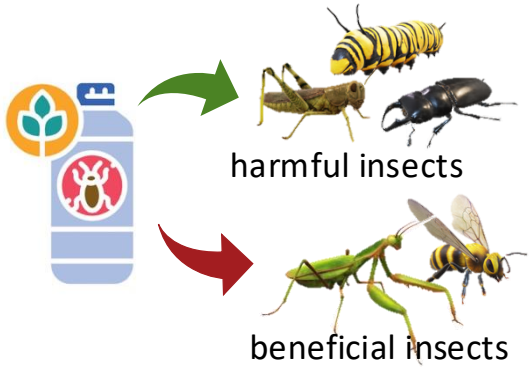
Increased global food demand

Effective crop protection

Why does off-target assessment of biopesticides matter?

Increased global food demand

Effective crop protection



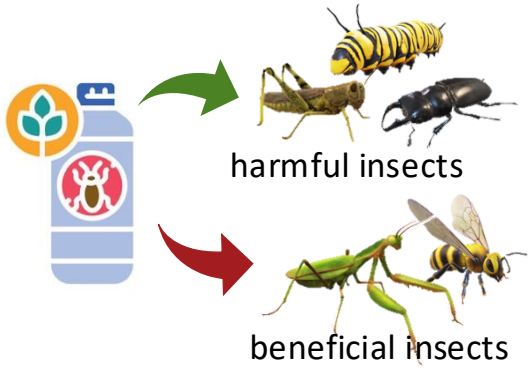
Novel biopesticides

Lack of EU regulations

Why does off-target assessment of biopesticides matter?

Increased global food demand

Effective crop protection



Novel biopesticides

Lack of EU regulations

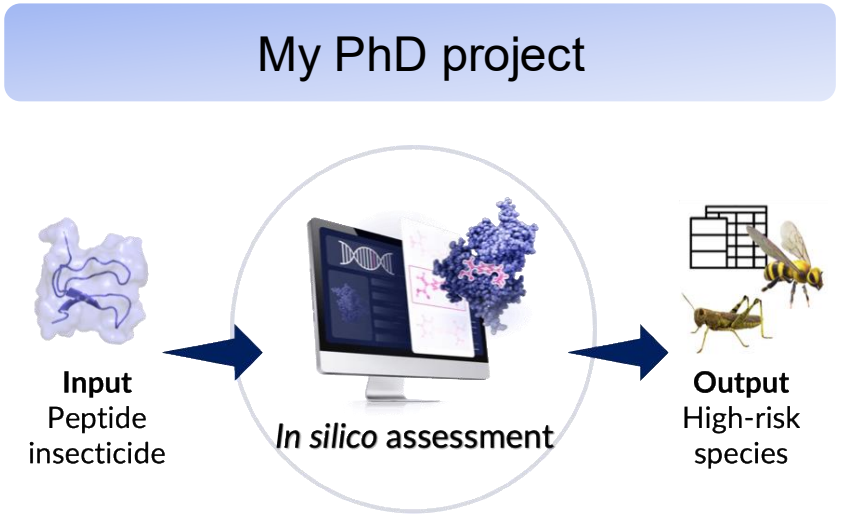
novo nordisk foundation
Novo Nordisk Foundation Challenge
Grant no. NNF24OC0087112

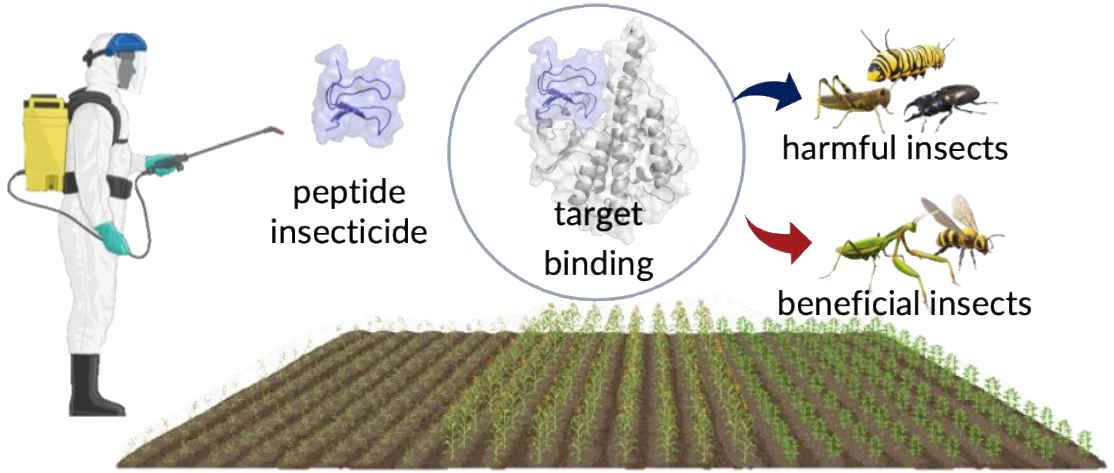
ENSAFE is an environmental risk assessment framework for siRNA- and peptide-based pesticides: integrating *in silico*, *in vitro*, and *in vivo* approaches.

Structural bioinformatics approaches and off-target binding prediction

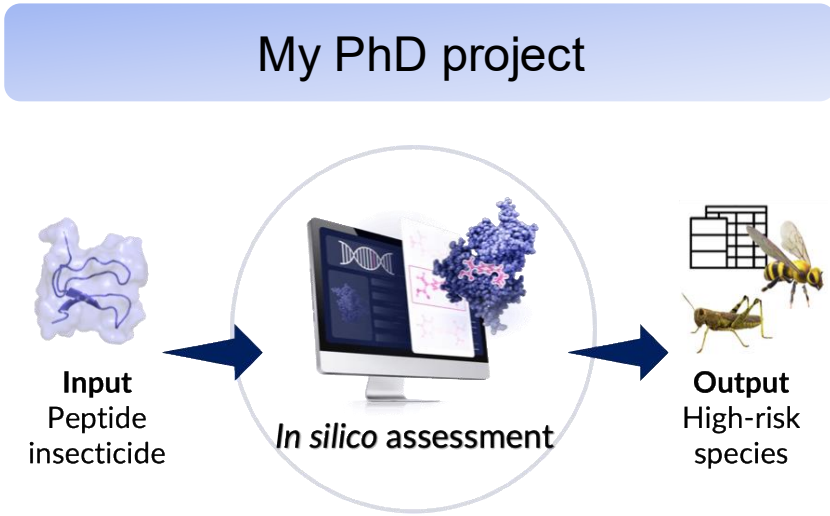
My PhD project

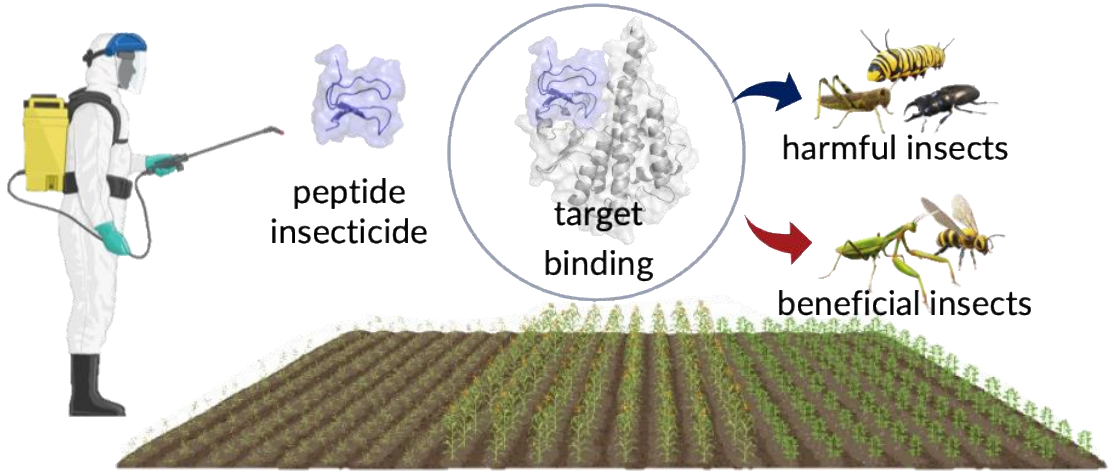
Structural bioinformatics approaches and off-target binding prediction





Structural bioinformatics approaches and off-target binding prediction





Hypothesis

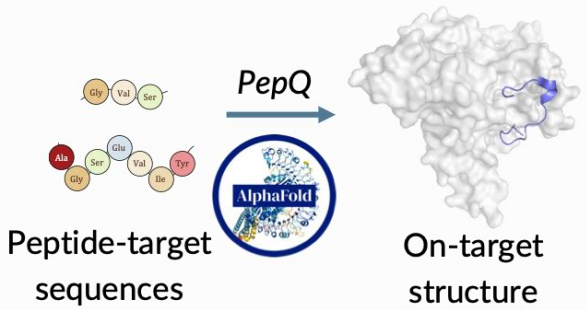
- Off-target effects are through promiscuous target binding, which can be predicted from on-target
  - Conserved binding motifs.
  - Off-target structural plausibility.

Off-target binding prediction through conserved binding motifs and structural plausibility

- Prediction workflow

Off-target binding prediction through conserved binding motifs and structural plausibility

- Prediction workflow

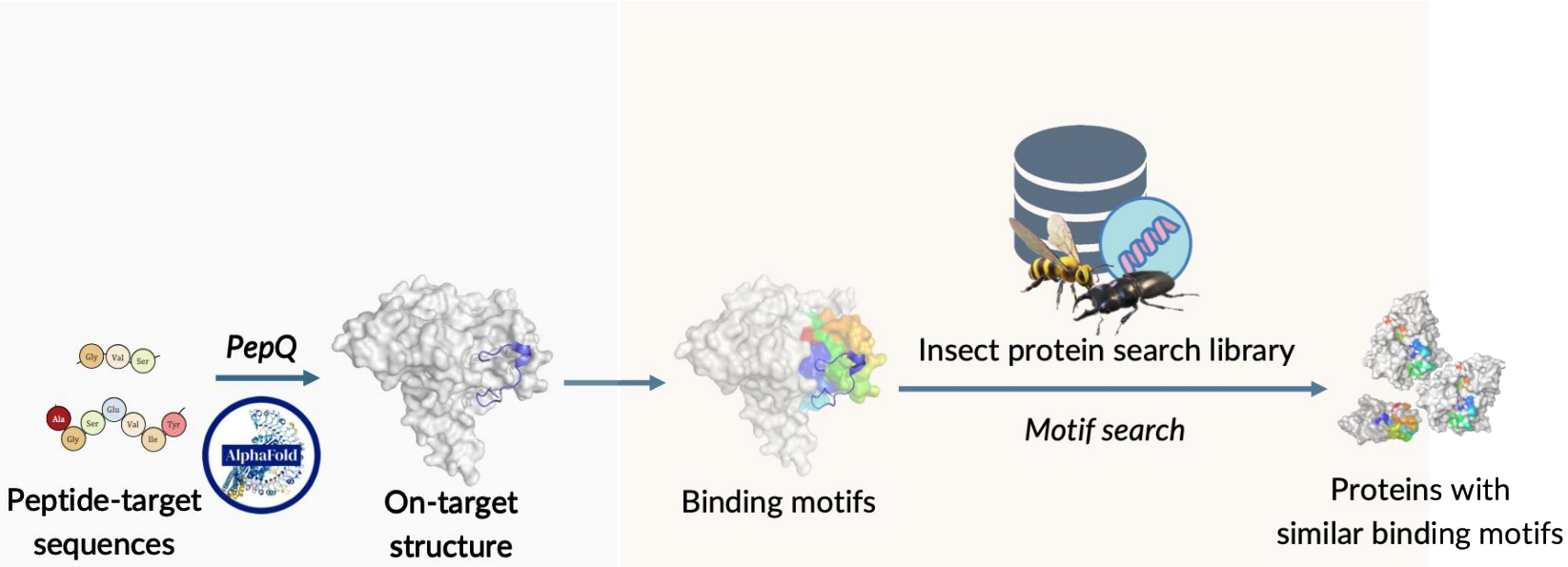


Module 1. On-target structure generation

Jumper, J *et al.* (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583-589.
Evans, R., *et al.* (2022). Protein complex prediction with AlphaFold-multimer. bioRxiv [Preprint](2022).

Off-target binding prediction through conserved binding motifs and structural plausibility

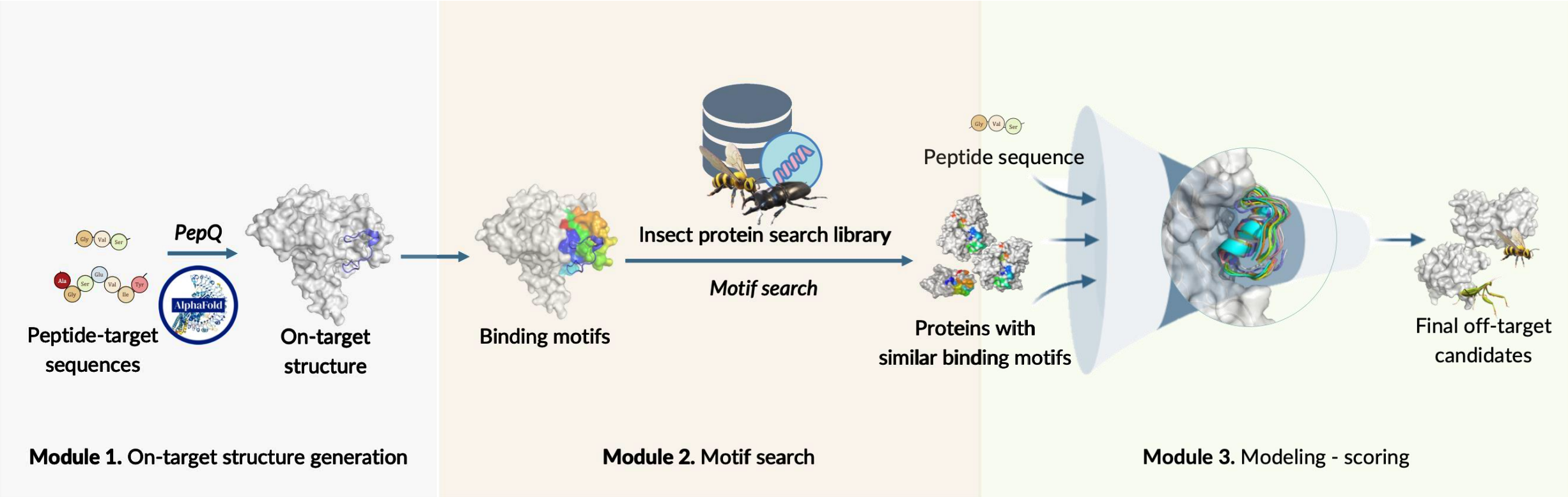
- Prediction workflow



Jumper, J *et al.* (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583-589.
Evans, R., *et al.* (2022). Protein complex prediction with AlphaFold-multimer. bioRxiv [Preprint](2022).

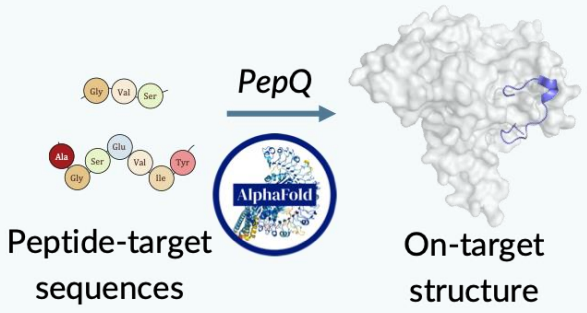
Off-target binding prediction through conserved binding motifs and structural plausibility

- Prediction workflow



Jumper, J *et al.* (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583-589.
Evans, R., *et al.* (2022). Protein complex prediction with AlphaFold-multimer. bioRxiv [Preprint](2022).

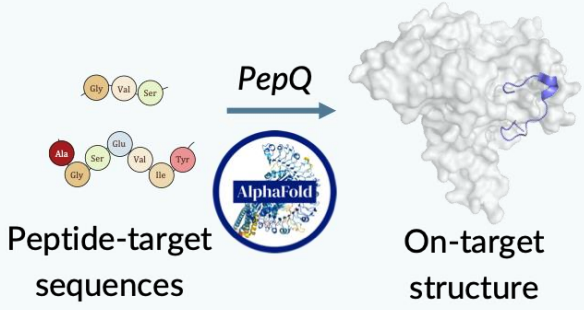
PepQ: peptide-protein model quality assessment



Module 1. On-target structure generation

- Generating the on-target structure generation using ColabFold
  - 5 models per run.
  - Confidence scores.

PepQ: peptide-protein model quality assessment

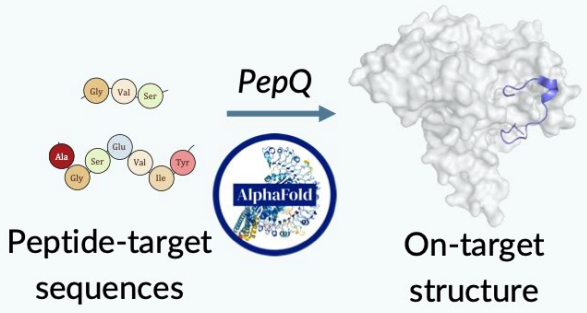


Module 1. On-target structure generation

- Generating the on-target structure generation using ColabFold
  - 5 models per run.
  - Confidence scores.

How to assess and select complexes with high quality?

PepQ: peptide-protein model quality assessment



Module 1. On-target structure generation

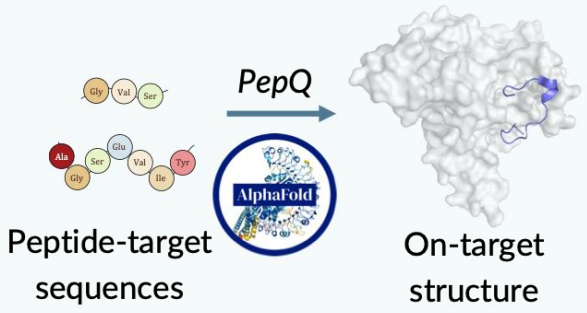
- Generating the on-target structure generation using ColabFold
  - 5 models per run.
  - Confidence scores.

Reference structure is present

DockQ

Similarities models vs. reference

PepQ: peptide-protein model quality assessment

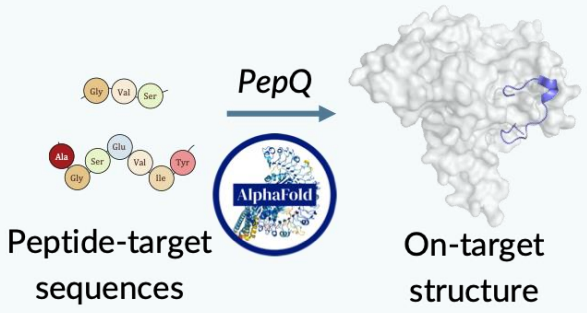


Module 1. On-target structure generation

- Generating the on-target structure generation using ColabFold
  - 5 models per run.
  - Confidence scores.

Reference structure is present	Reference structure is absent
DockQ Similarities models vs. reference	**pdockq, pdockq2** Predicted DockQ for protein-protein models

PepQ: peptide-protein model quality assessment



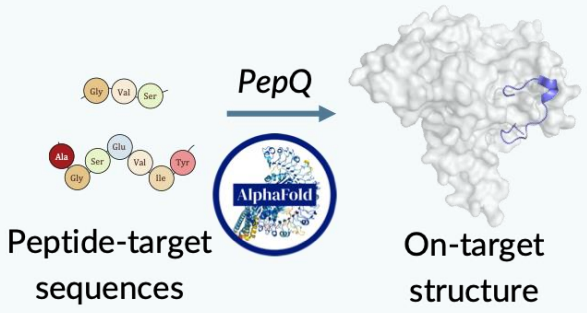
Module 1. On-target structure generation

- Generating the on-target structure generation using ColabFold
  - 5 models per run.
  - Confidence scores.

Reference structure is present	Reference structure is absent
DockQ Similarities models vs. reference	**pdockq, pdockq2** Predicted DockQ for protein-protein models

No available quality assessment for peptide-protein models

PepQ: peptide-protein model quality assessment



Module 1. On-target structure generation

- Generating the on-target structure generation using ColabFold
  - 5 models per run.
  - Confidence scores.

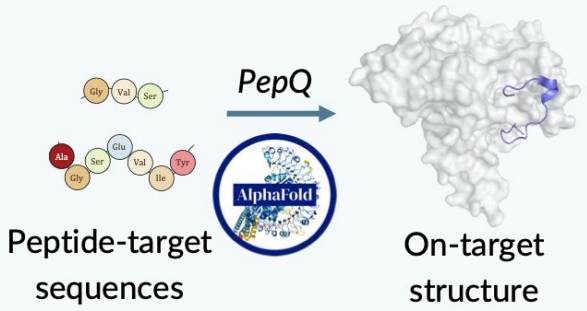
Reference structure is present	Reference structure is absent
DockQ Similarities models vs. reference	**pdockq, pdockq2** Predicted DockQ for protein-protein models

No available quality assessment for peptide-protein models

PepQ

Predicted DockQ for peptide-protein models

PepQ: peptide-protein model quality assessment



Module 1. On-target structure generation

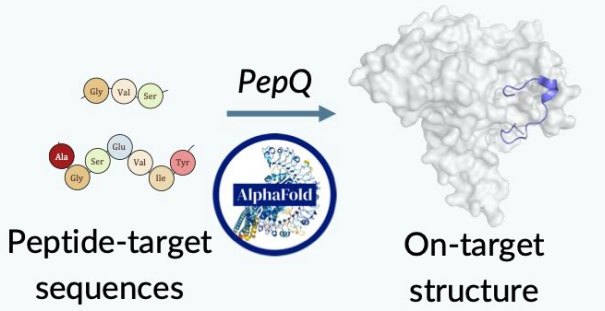
Data

Query peptide-protein complexes from Protein Data Bank with the criteria regarding resolution, canonical amino acids, lengths, and redundancy filter.

- Training set: complexes released before the release date of AlphaFold-Multimer.
- Test set: complexes released after the release date of AlphaFold-Multimer.
- Similarity filtering: any complexes in the test set that are similar to those in the training set were removed

Bryant, P. *et al.* (2022). Improved prediction of protein-protein interactions using AlphaFold2. Nature communications, 13(1), 1265.
Zhu, W. *et al.* (2023). Evaluation of AlphaFold-Multimer prediction on multi-chain protein complexes. *Bioinformatics*, *39*(7), btad424.

PepQ: peptide-protein model quality assessment



Module 1. On-target structure generation

Data

Query peptide-protein complexes from Protein Data Bank with the criteria regarding resolution, canonical amino acids, lengths, and redundancy filter.

- Training set: complexes released before the release date of AlphaFold-Multimer.
- Test set: complexes released after the release date of AlphaFold-Multimer.
- Similarity filtering: any complexes in the test set that are similar to those in the training set were removed

Competing scores

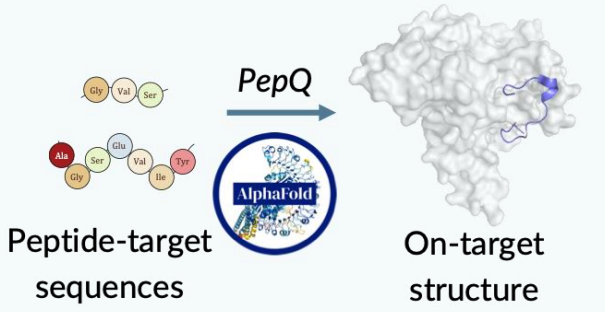
- pdockq: predicted protein-protein complex quality score (Bryant, 2022)
- pdockq2: predicted protein-protein complex quality score (Zhu, 2023)
- AlphaFold-Multimer internal ranking scores.

Model architecture

- Multi-head attention: regression head (Huber), ranking head (ranking loss)

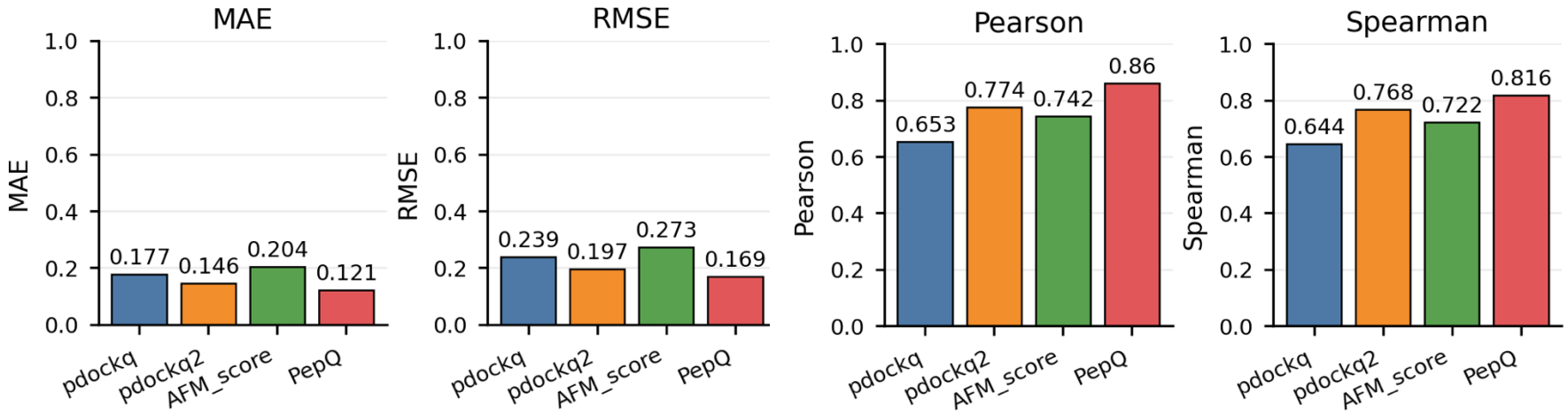
Bryant, P. *et al.* (2022). Improved prediction of protein-protein interactions using AlphaFold2. Nature communications, 13(1), 1265.
Zhu, W. *et al.* (2023). Evaluation of AlphaFold-Multimer prediction on multi-chain protein complexes. *Bioinformatics*, *39*(7), btad424.

PepQ: peptide-protein model quality assessment



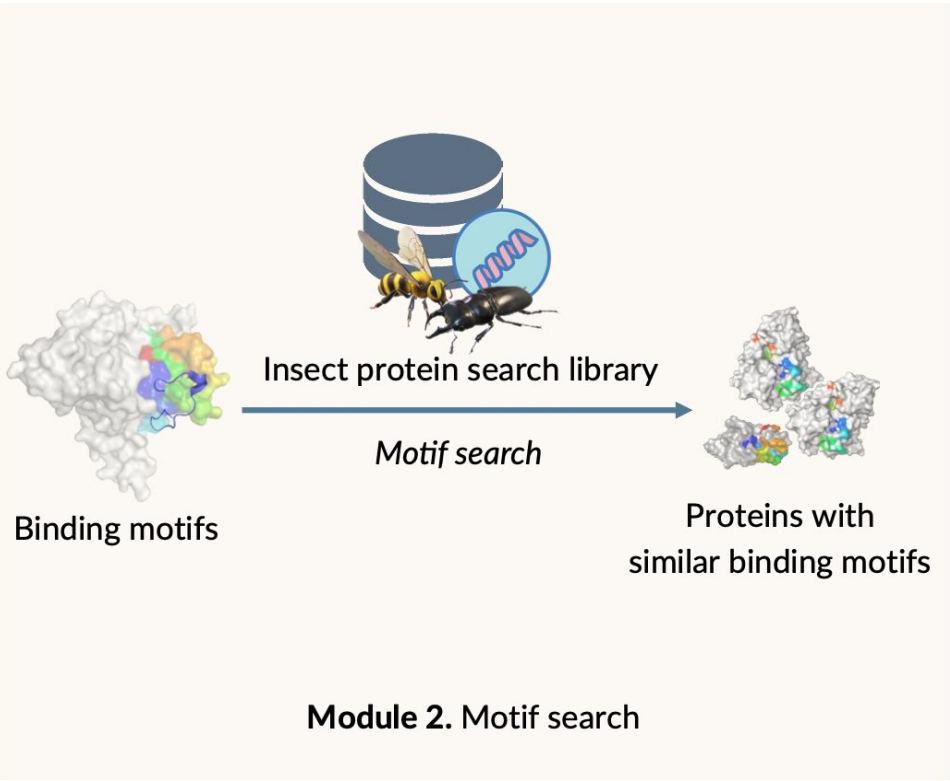
Module 1. On-target structure generation

Performance comparison on test set



- pdockQ and pdockQ2 are logarithmic regression models
  - pdockQ relies on inter-chain contact counts and interface pLDDT.
  - pdockQ2 enhances this framework by incorporating PAE.
- PepQ further integrates peptide-specific confidence features, allowing more accurate assessment of peptide-protein interfaces.

Structure-based motif search



Module 2. Motif search

Search tool

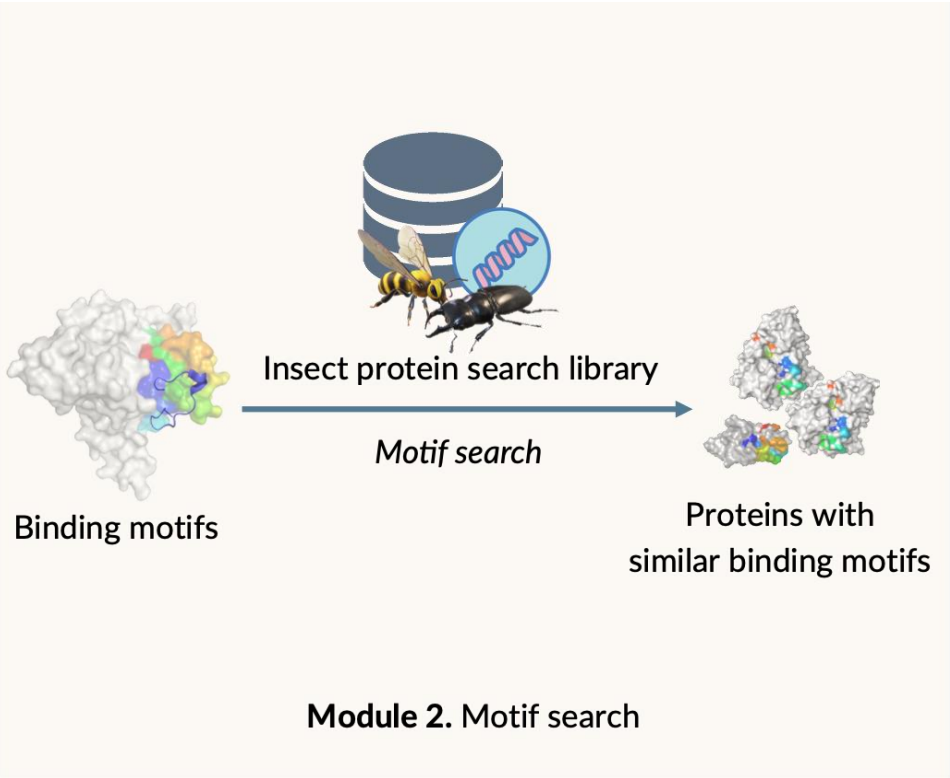
- Folddisco
  - Amino-acid identity, interatomic distances, and backbone/side-chain angular features.
  - Fast lookups against large structural databases

Protein library

- AlphaFold Protein Structure Database
- Indexed

Kim, H., *et al.* (2025). Structural motif search across the protein-universe with Folddisco. *bioRxiv*, 2025-07.

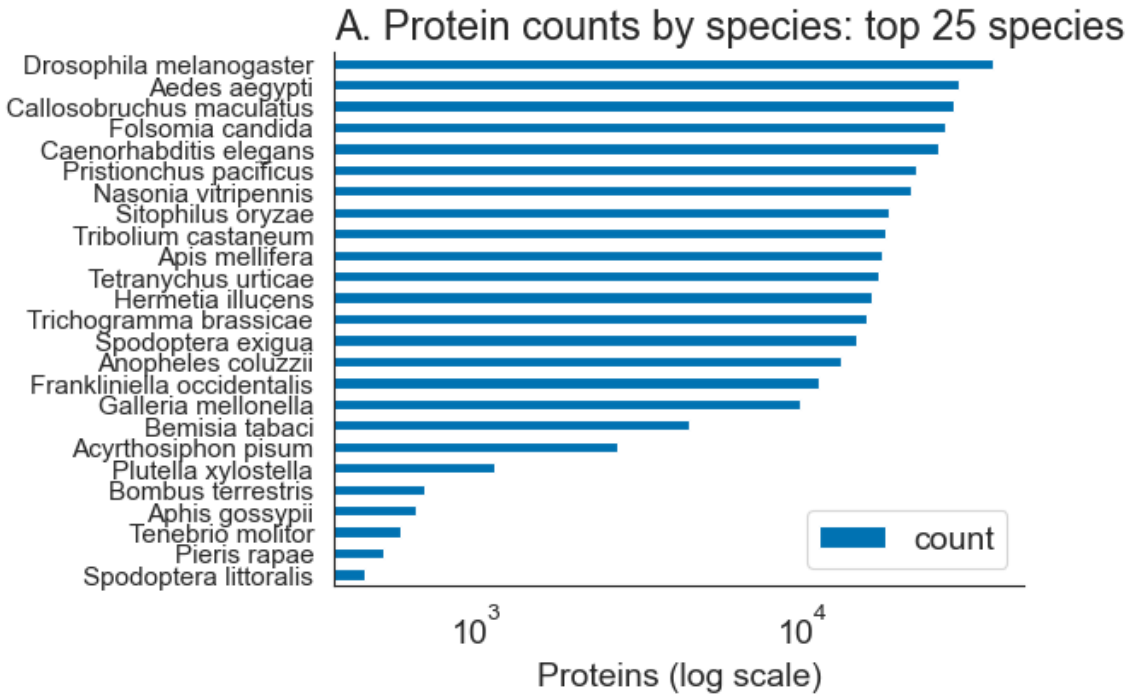
Protein library



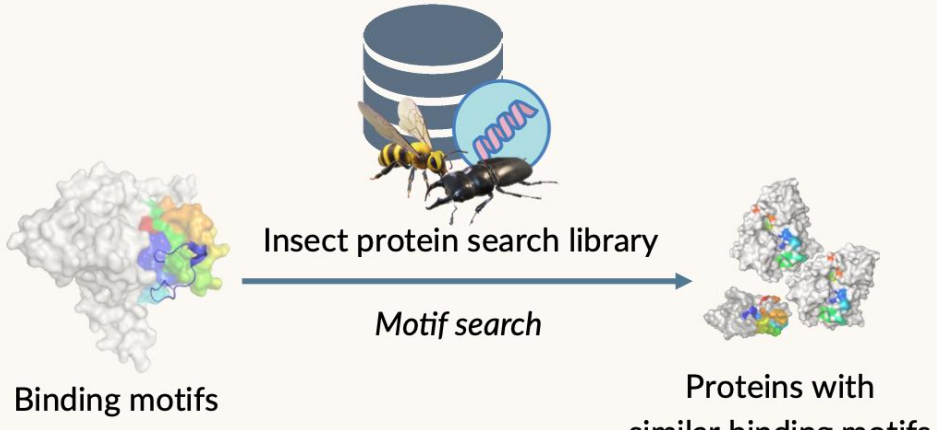
Module 2. Motif search

Pre-indexed, ready-to-use library

- 367 142 proteins
- 99 species of interest



Future steps

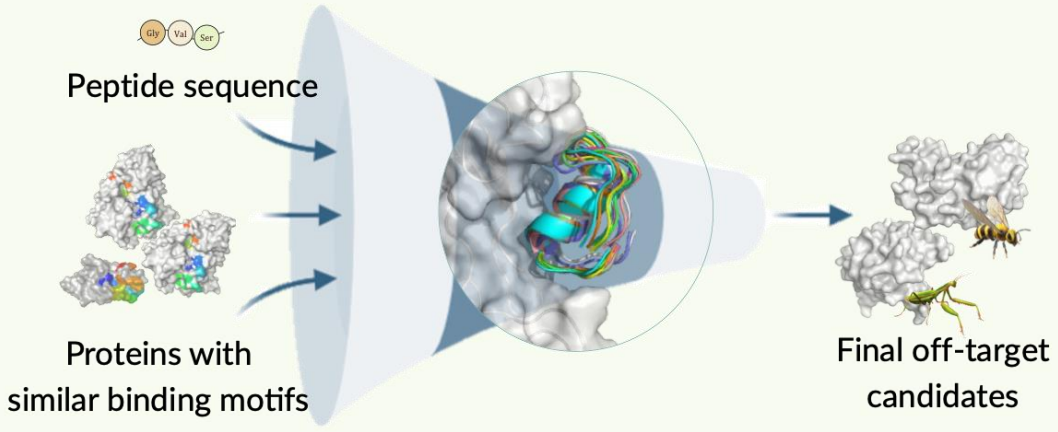


Module 2. Motif search

Module 2. Motif search

- Improve quantity and quality of protein library.

Future steps



Module 3. Modeling - scoring

Module 2. Motif search

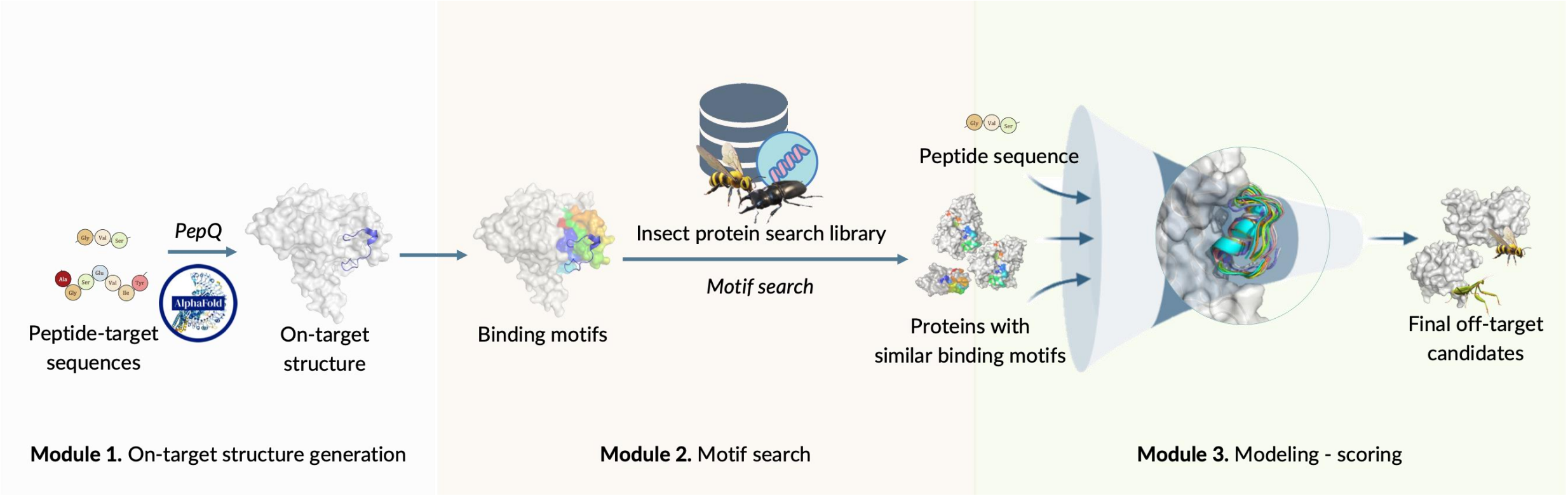
- Improve quantity and quality of protein library.

Module 3. Modelling - scoring

- Curate and generate training datasets reflecting off-target binding.
- Develop models to predict off-target binding probability.

Off-target binding prediction through conserved binding motifs and structural plausibility

- Prediction workflow



Acknowledgements

- **RTH**
  - **Christian Anthon**
  - Lorenzo Favaro
  - Buwen Liang
  - Nadezhda Tsankova Doncheva
  - Ying Sun
  - Mikkel Frier Wolfhagen
  - Betina Wingreen Jensen
  - Jakob Hull Havgaard
  - Stefan E. Seemann
  - Stefano Roncelli
  - Gül S. Demircan
  - **Jan Gorodkin**
- **Collaborators**
  - Arne Elofsson, Science for Life Laboratory, Stockholm University, Sweden
  - ENSAFE consortium
- **Funding**
  - Novo Nordisk Foundation

RTH CENTER FOR NON-CODING RNA IN TECHNOLOGY AND HEALTH

UNIVERSITY OF COPENHAGEN

novo nordisk foundation
Novo Nordisk Foundation Challenge

ENSAFE

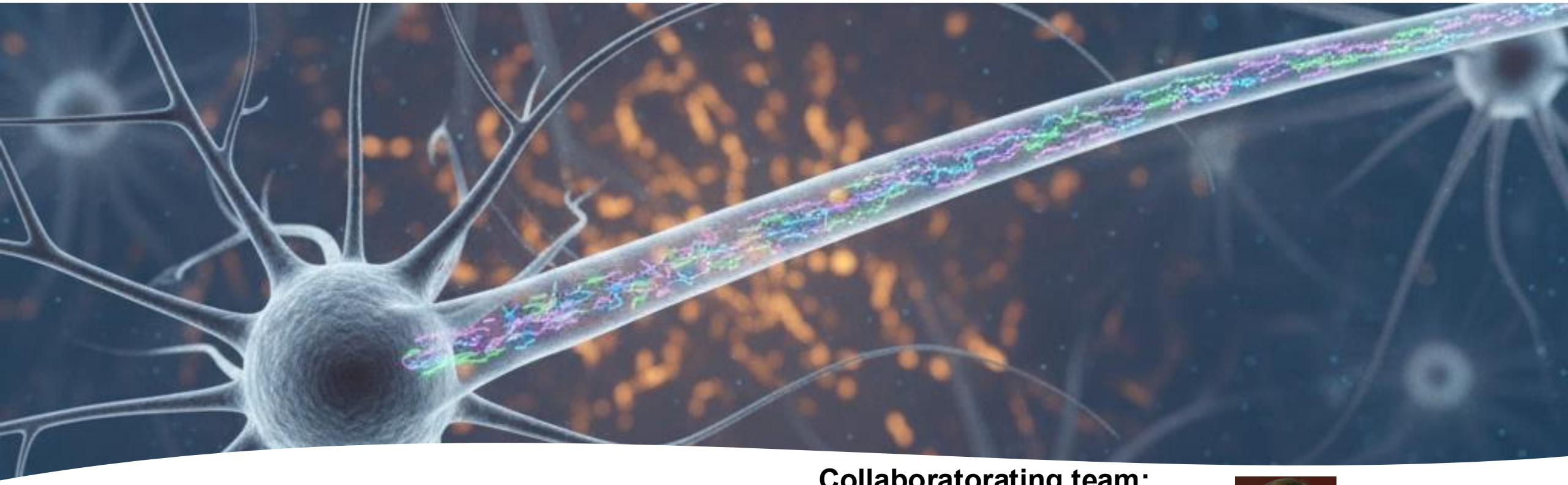
Thank you for your attention!

RTH
CENTER FOR NON-CODING RNA IN TECHNOLOGY AND HEALTH

ENSAFE

UNIVERSITY OF COPENHAGEN

Upcoming PhD or postdoc position at RTH
Talk to Jan

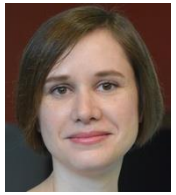


Decoding RNA journey:

Predictive models for the subcellular RNA fate in neurons

Collaboratorating team:

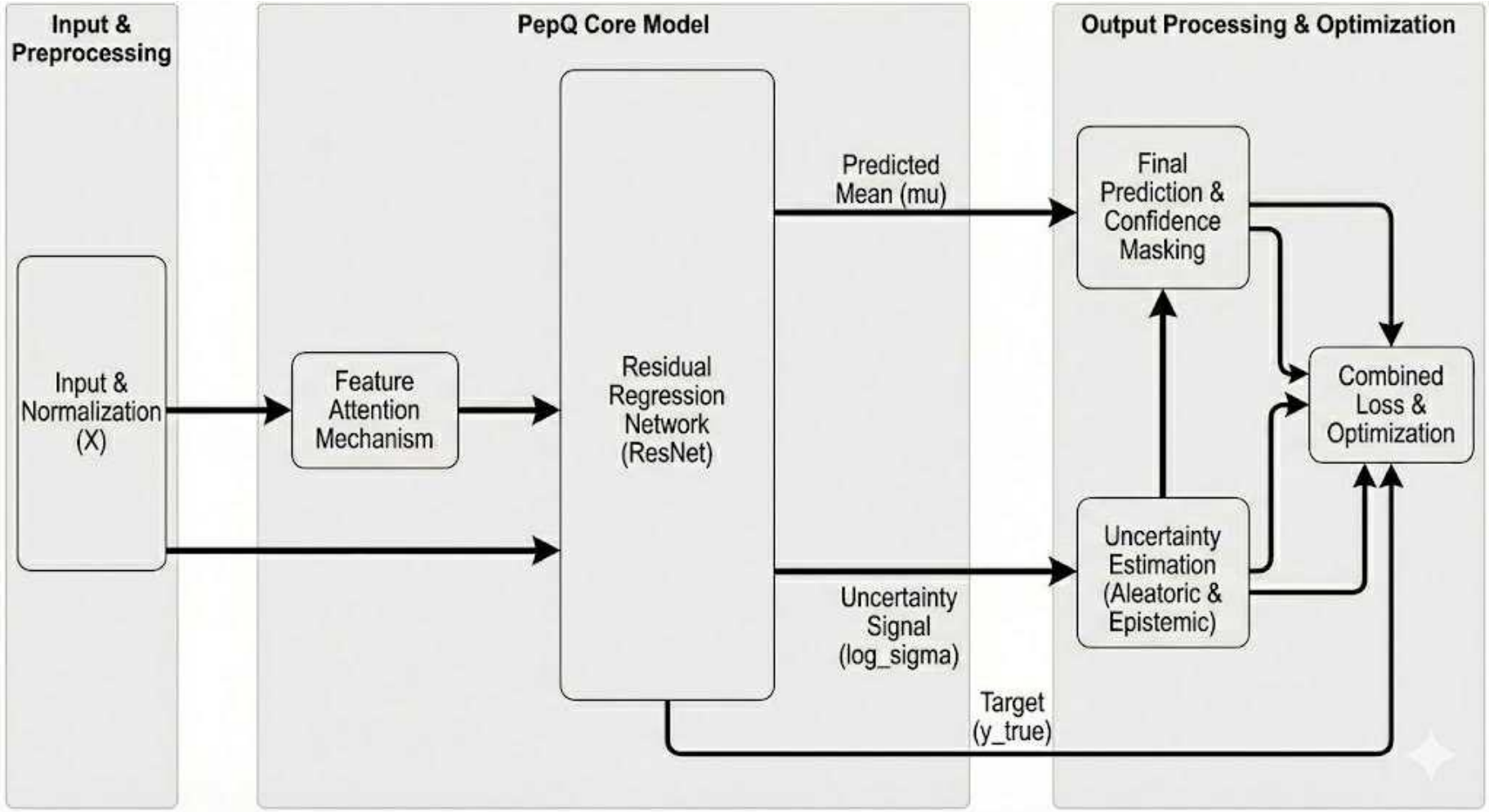
Chekulaeva (UCPH / Berlin)



Ulitsky (Weizmann Institute)

Gorodkin (UCPH)

PepQ: peptide-protein model quality assessment



PepQ: peptide-protein model quality assessment

Full Objective

$$L(\theta; e) = w_{reg}\ \mathcal{L}_{reg} + w_{rank}(e)\,\mathcal{L}_{rank} + w_{nll}\ \mathcal{L}_{nll} + w_{attn\_entropy}\,\mathcal{L}_{ent} + w_{gate\_l2}\,\mathcal{L}_{gate}$$

Regression Loss: $\mathcal{L}_{reg} = \sum_{i=1}^{B} (y_i - \mu_i)^2$ *or* Huber Loss

Ranking Loss: $\mathcal{L}_{rank} = w_{rank}(e) \sum_{i,j \in P}^{P} w_{ij} \log\left(1 + \exp\left(-s_{ij}(\Delta\mu_{ij} - m)\right)\right).$

Gaussian NLL Loss: $\mathcal{L}_{nll} = \sum_{i=1}^{B} \left(\log \sigma_i + \frac{1}{2}\left(y_i - \mu_i\right)^2\right),$

Attention Entropy Loss: $\mathcal{L}_{ent} = w_{attn\_entropy}\, H = w_{attn\_entropy} \frac{H}{1}{BH} \sum_{i=1}^{B} \sum_{H=1} \sum_{f=1}^{H} - a_{i,h,f} \log a_{i,h,f}$

Gate L2 Loss: $\mathcal{L}_{gate} = w_{gate\_l2} \sum_{f=D}^{B} \left(g_{i,f} - 1\right))^2.$