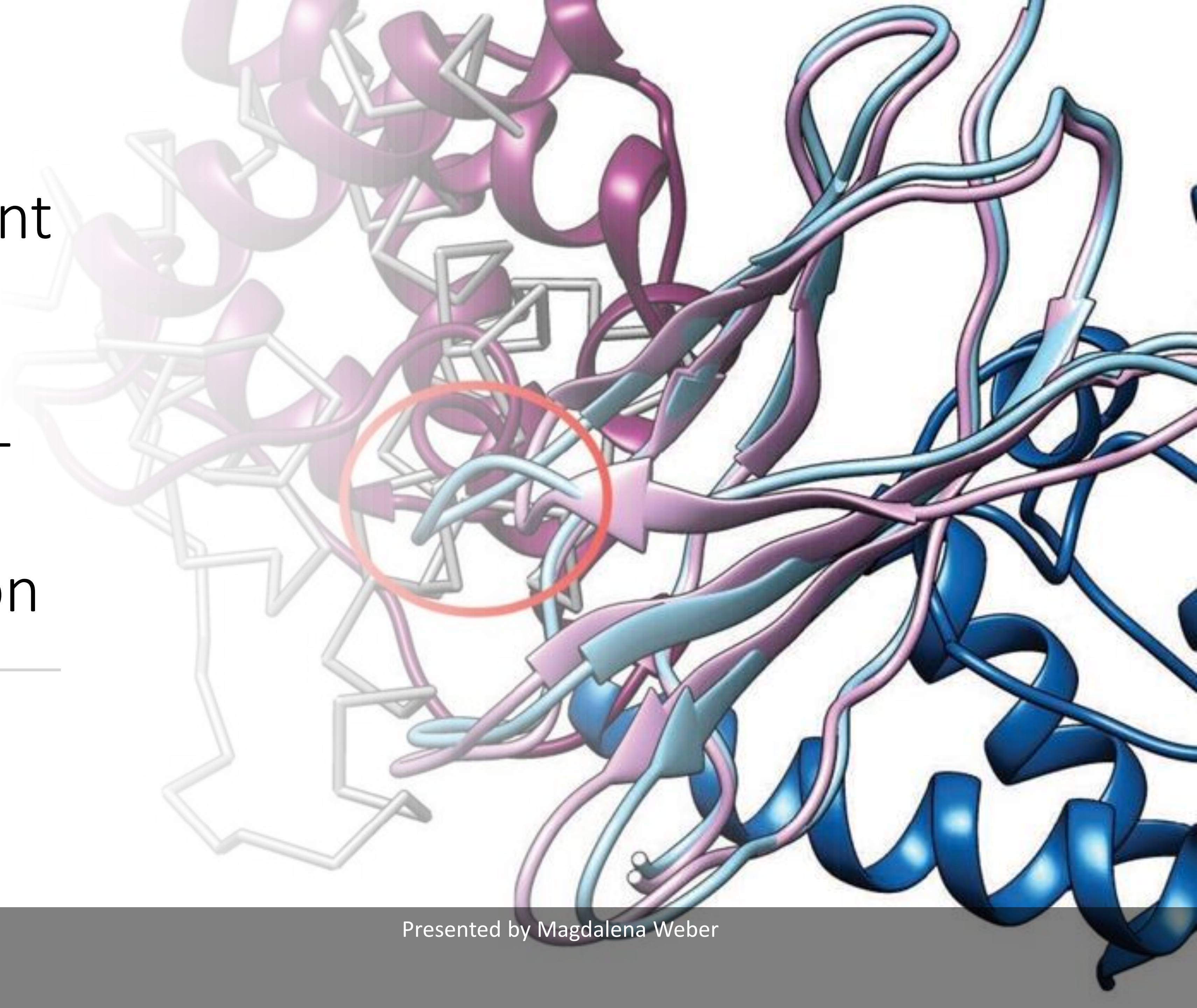
Sequence alignment using machine learning for accurate template-based protein structure prediction

Shuichiro Makigaki,

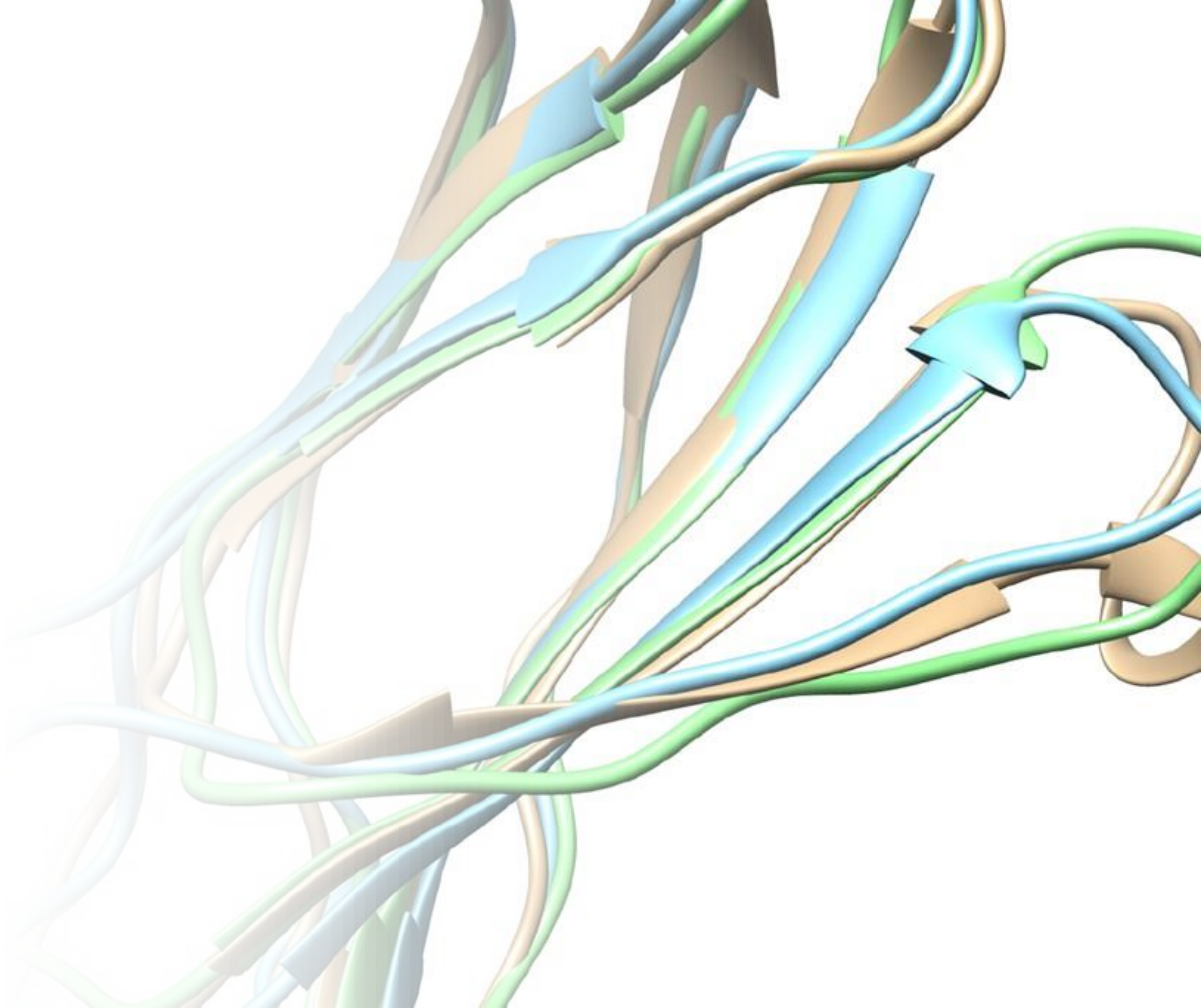
Takashi Ishida



Presented by Magdalena Weber

Introduction

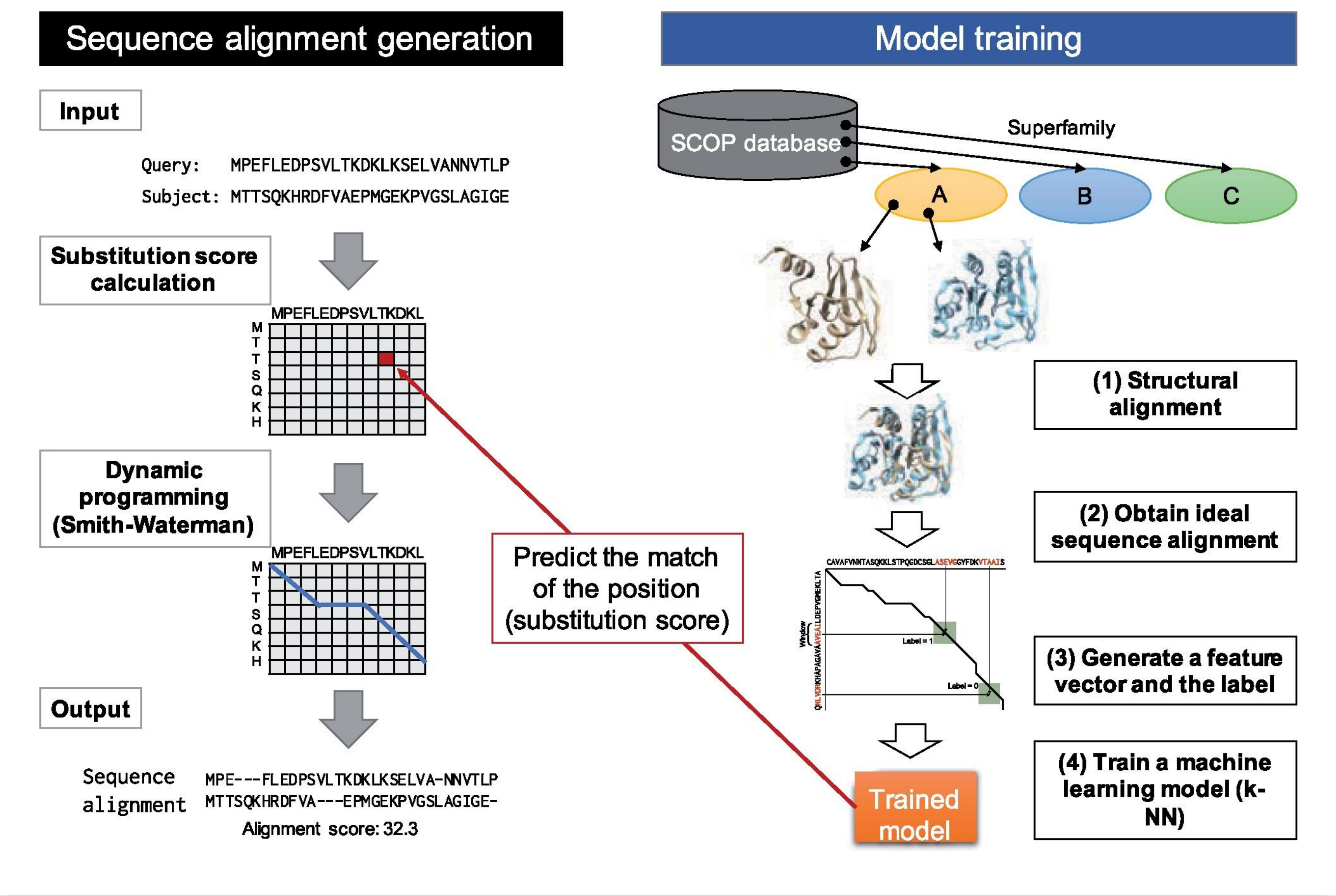
- What is template-based modeling (TBM)?
- Why do we need TBM?
- How does it work?

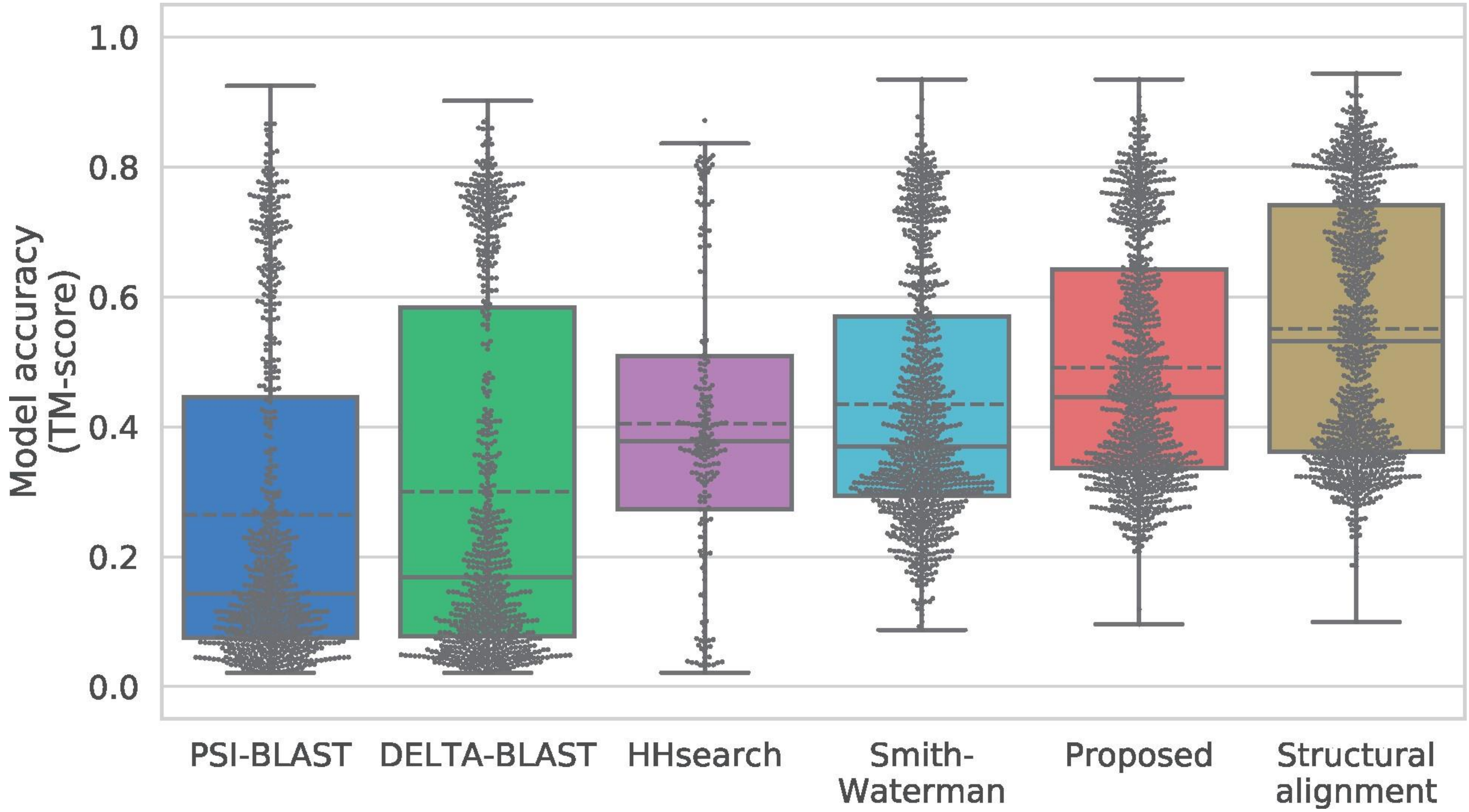


Motivation

- The need for optimal method for protein structure prediction using sequence data
- Improving TBM
  - → Shift focus to accuracy
    - →Alignment quality is crucial
      - →Focus on improving alignments

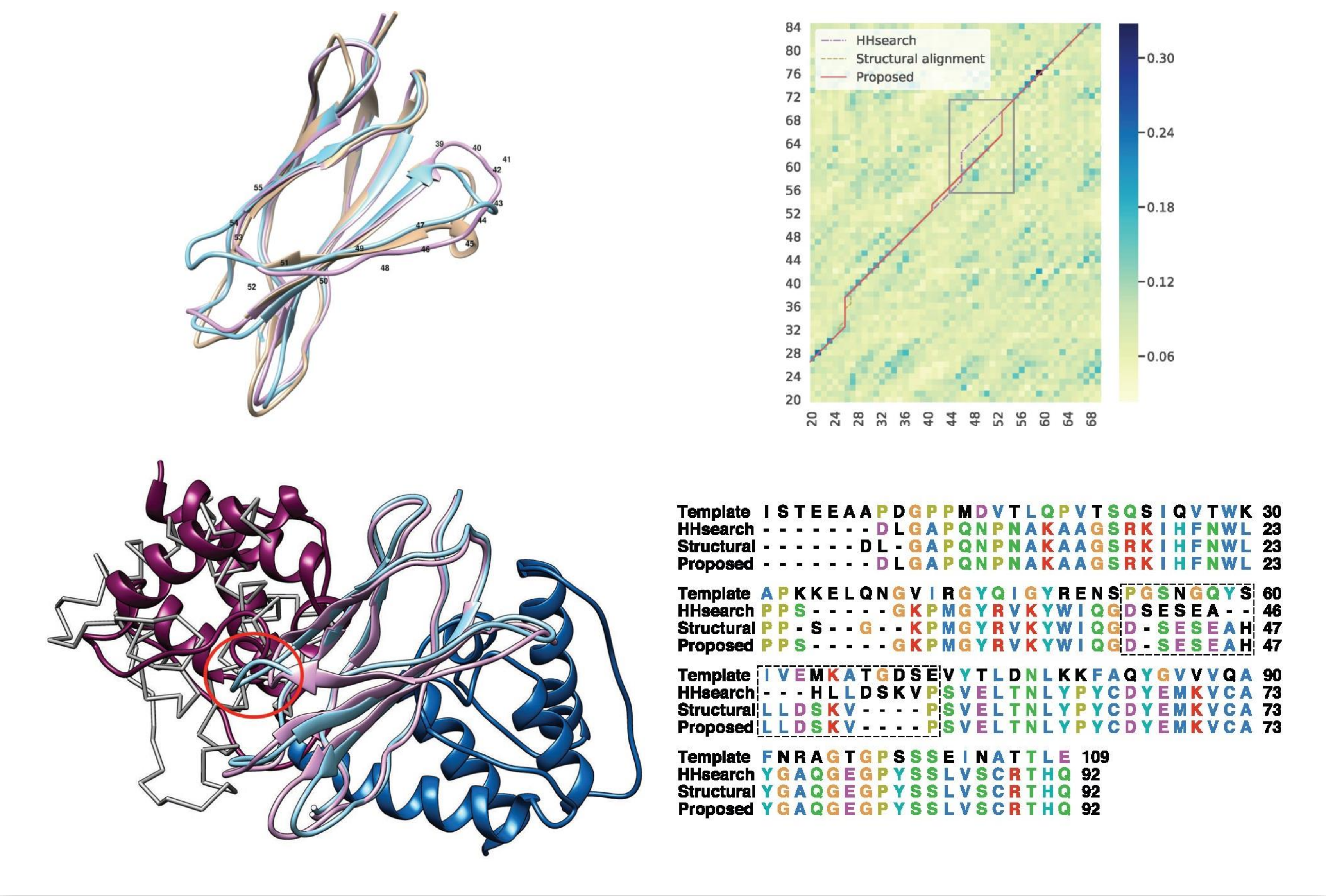
Methods


Sequence alignment generation
Input
Query: MPEFLEDPSVLTKDKLKSELVANNVTLP
Subject: MTTSQKHRDFVAEPMGEKPVGSLAGIGE
Substitution score calculation
MPEFLEDPSVLTKDKL
M T T S Q K H
Dynamic programming (Smith-Waterman)
MPEFLEDPSVLTKDKL
M T T S Q K H
Output
Sequence alignment
MPE---FLEDPSVLTKDKLKSELVA-NNVTLP
MTTSQKHRDFVA---EPMGEKPVGSLAGIGE-
Alignment score: 32.3
Predict the match of the position (substitution score)
Model training
SCOP database
Superfamily
A
B
C
(1) Structural alignment
(2) Obtain ideal sequence alignment
CAVAFVNNTASQKKLSTPQGDCSGLASEVGGYFDKVTAAIS
QNLVDRKHAPAGAVAAVEAILLDEPVGMEKLTA
Window
Label = 1
Label = 0
(3) Generate a feature vector and the label
(4) Train a machine learning model (k-NN)
Trained model


1.0
0.8
0.6
0.4
0.2
0.0
Model accuracy (TM-score)
PSI-BLAST
DELTA-BLAST
HHsearch
Smith-Waterman
Proposed
Structural alignment

Results

Results



```
Template   I S T E E A A P D G P P M D V T L Q P V T S Q S I Q V T W K 30
HHsearch   - - - - - - - D L G A P Q N P N A K A A G S R K I H F N W L 23
Structural - - - - - - D L - G A P Q N P N A K A A G S R K I H F N W L 23
Proposed   - - - - - - - D L G A P Q N P N A K A A G S R K I H F N W L 23

Template   A P K K E L Q N G V I R G Y Q I G Y R E N S P G S N G Q Y S 60
HHsearch   P P S - - - - - G K P M G Y R V K Y W I Q G D S E S E A - - 46
Structural P P - S - - G - - K P M G Y R V K Y W I Q G D - S E S E A H 47
Proposed   P P S - - - - - G K P M G Y R V K Y W I Q G D - S E S E A H 47

Template   I V E M K A T G D S E V Y T L D N L K K F A Q Y G V V V Q A 90
HHsearch   - - - H L L D S K V P S V E L T N L Y P Y C D Y E M K V C A 73
Structural L L D S K V - - - - P S V E L T N L Y P Y C D Y E M K V C A 73
Proposed   L L D S K V - - - - P S V E L T N L Y P Y C D Y E M K V C A 73

Template   F N R A G T G P S S S E I N A T T L E 109
HHsearch   Y G A Q G E G P Y S S L V S C R T H Q 92
Structural Y G A Q G E G P Y S S L V S C R T H Q 92
Proposed   Y G A Q G E G P Y S S L V S C R T H Q 92
```

Outlook

- Improving the proposed method
  → Reduce execution time
- Use of higher performance models like CNN
- Development of advanced methods of protein structure prediction

The End

Thank you for your attention