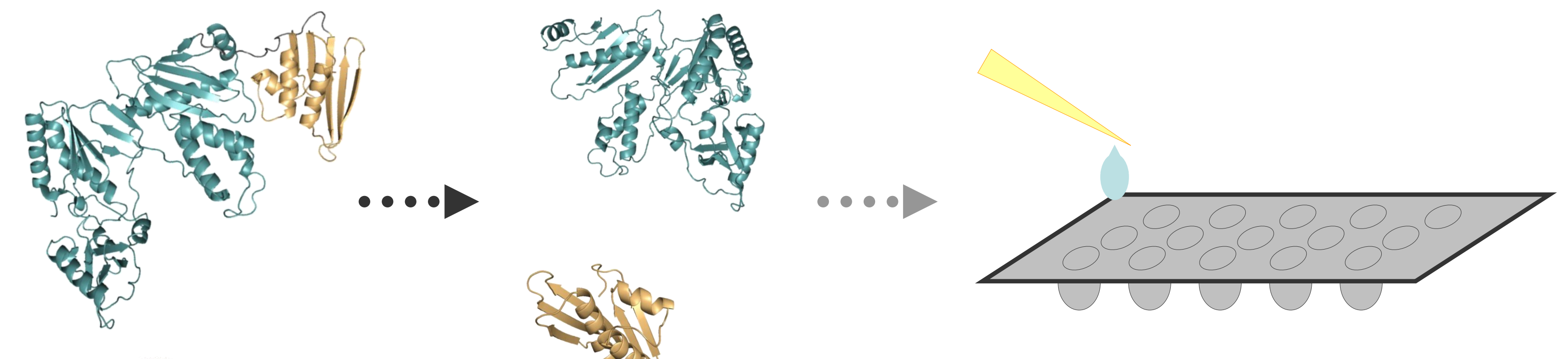
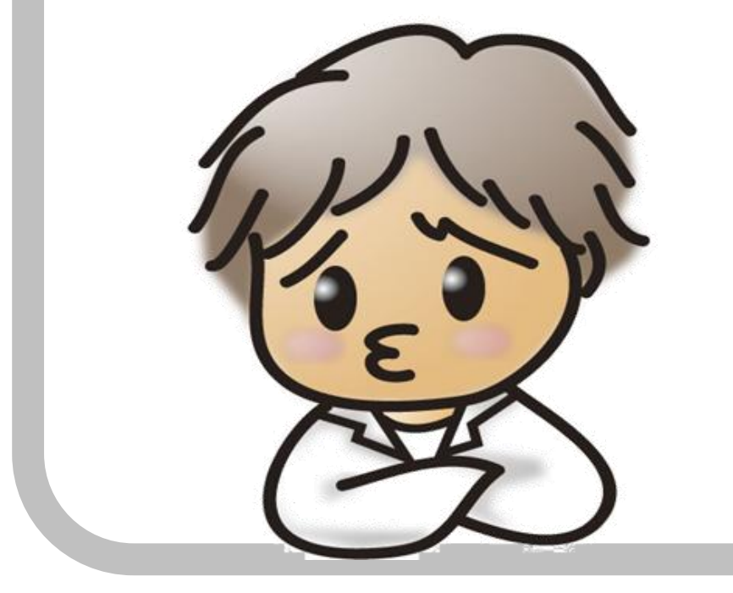
An SVM-Domain Linker Prediction Trained with Optimized Features Selected by Random Forest and Stepwise Selection

Teppei Ebina[1], Hiroyuki Toh[2] and Yutaka Kuroda[1]

1 Department of Biotechnology and Life Science, Tokyo University of Agriculture and Technology.
2 Medical Institute of Bioregulation, Kyushu University.

Introduction

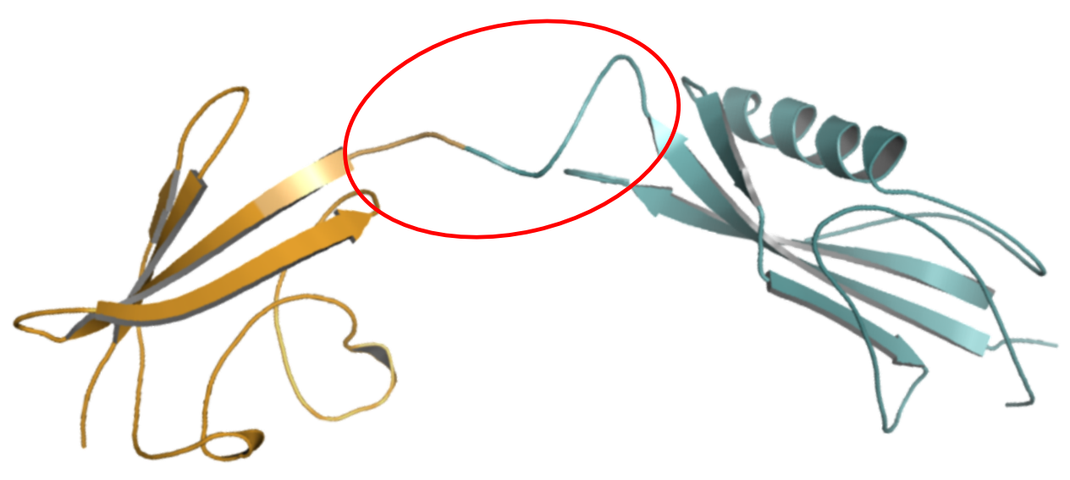




The prediction of structural domains has practical implication because large proteins often need to be dissected into structurally independent domains, which are usually easier to express, purify and characterize than whole proteins. Our specific goal is to develop an accurate domain linker prediction method & improve their prediction performances.

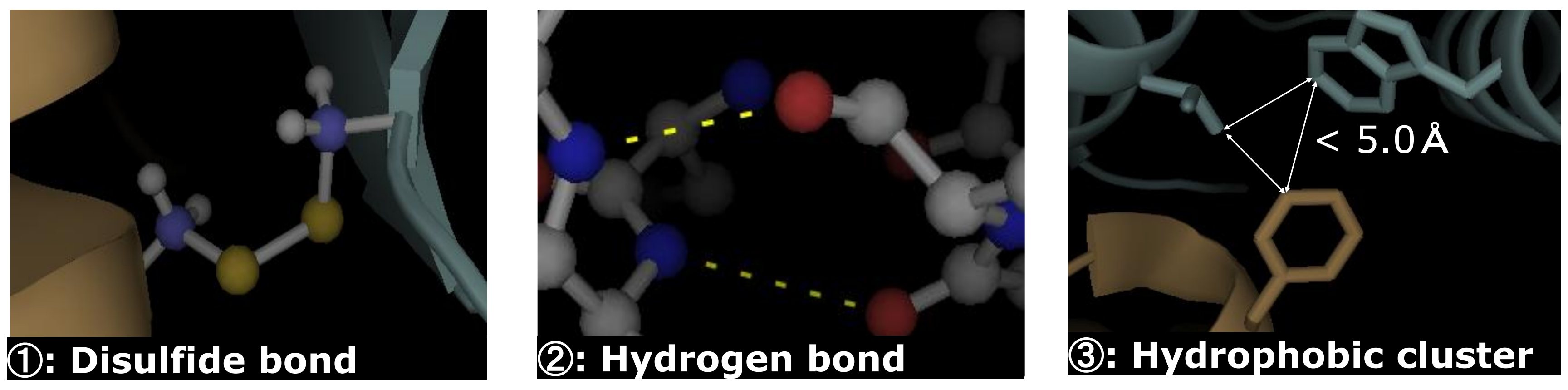
Methods

Target: Domain linker



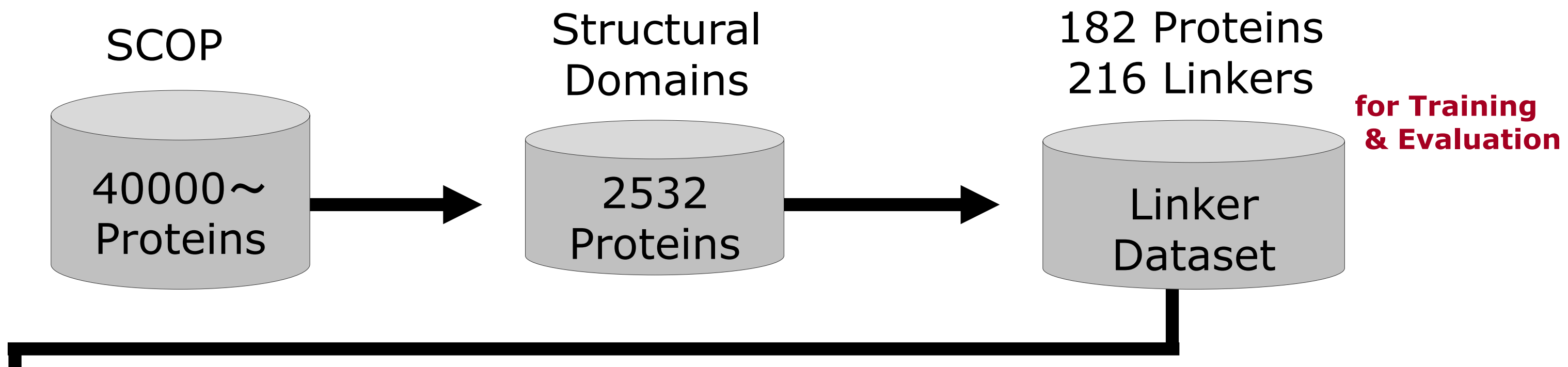
- Loop regions between two structural domains
- Easier to predict than domain regions

Structural domain



Domains having no inter-domain interactions

Predictor construction



Features

544 Amino Acid Indices
PSSM Elements
Probability of Secondary Structure
α-Helix & β-Sheet Core
Sequence Hydrophobic Core
Sequence Complexity
Similarity in Amino Acid Composition
between Domain
between Linker
Ratio of the Similarity Scores

Vector Coding

1, 11, 21, 31 or 41 residue window

···DTO···FHFFKQNVM···

Target Sequences

Vector Data

2870 dimensional vectors

Feature Selection – 1st Step

Random Forest Classification

Random Selection
200 domain & 200 linker data

Perform 100 times

Random Forest

Features with Z-Score of MDGI > 2.0 were selected as optimal feature candidates.

47 Optimal Feature Candidates

Feature Selection – 2nd Step

Backward Selection

Repeat until no improvement was observed by eliminating features

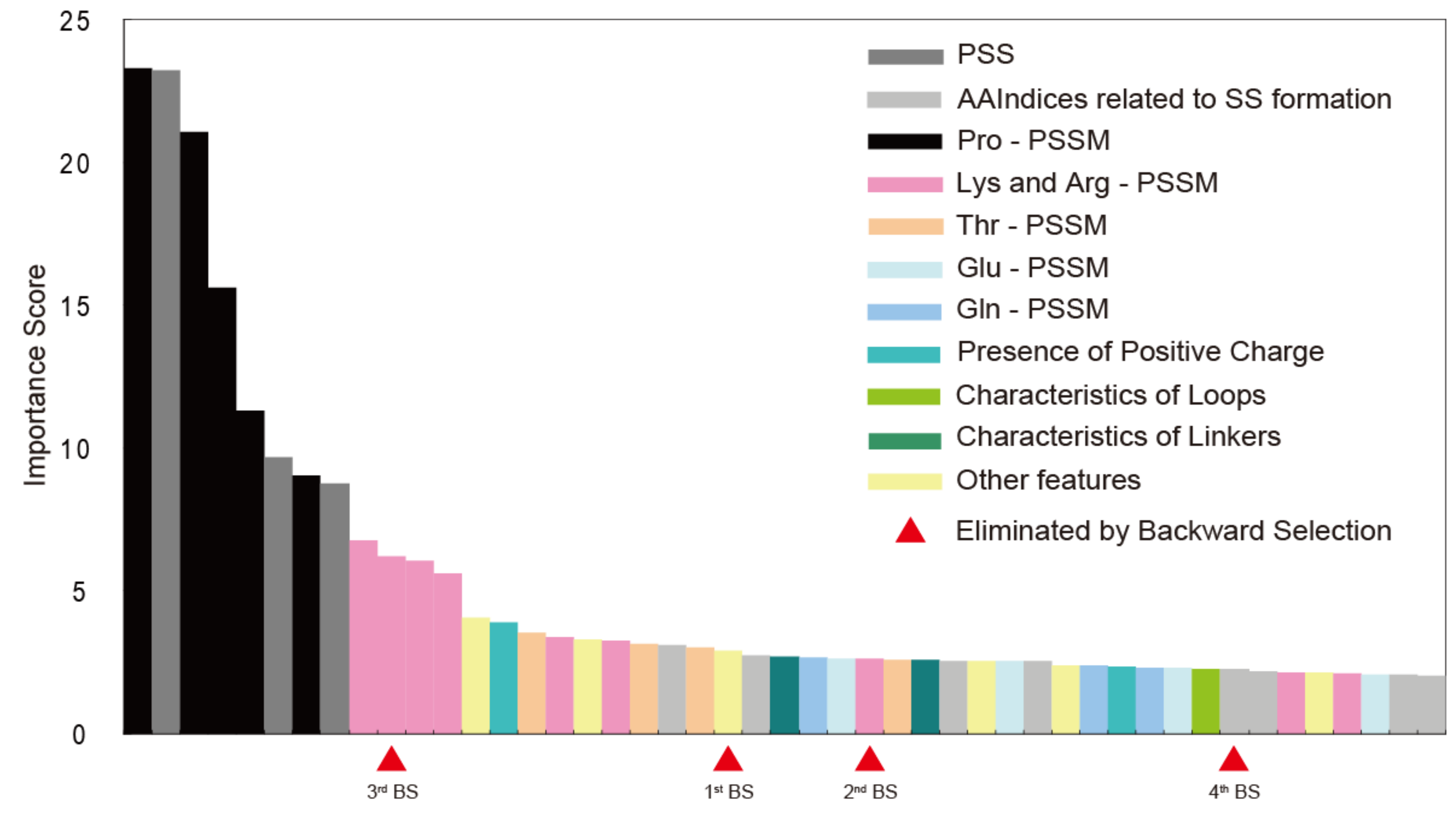
SVM

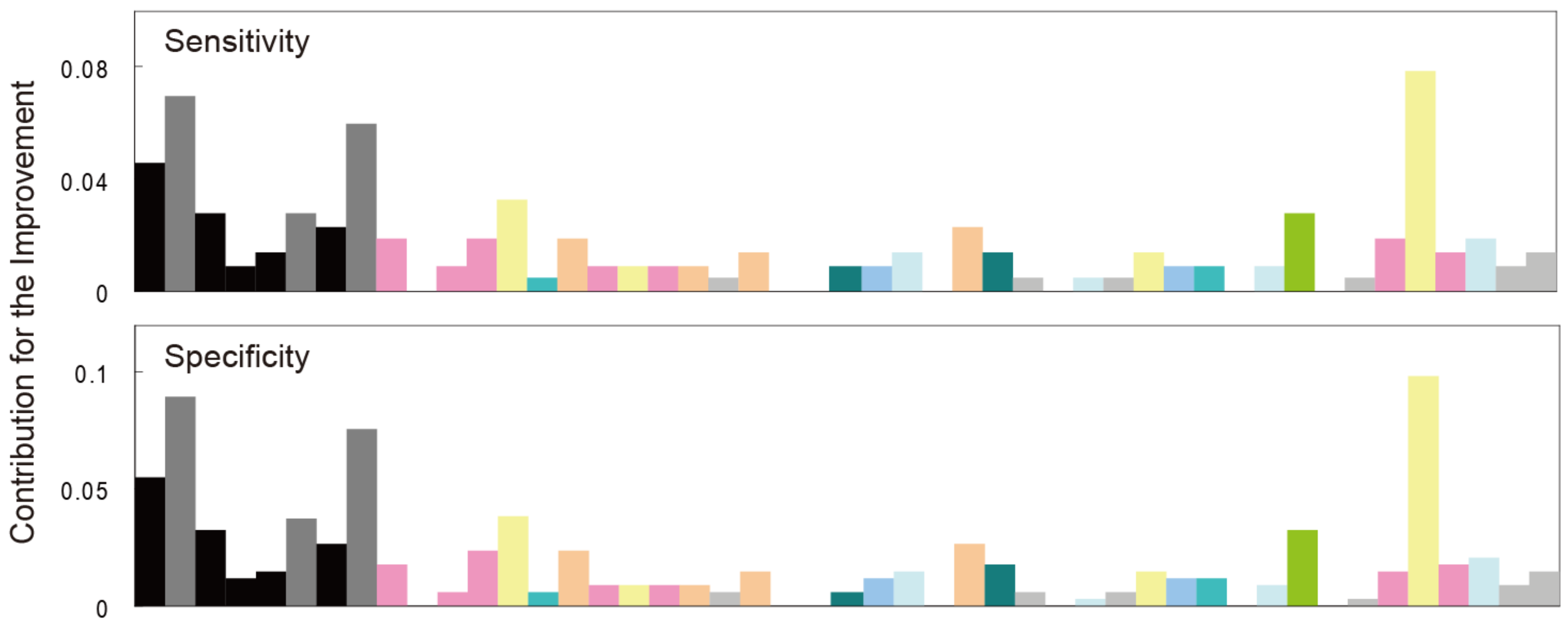
Assess the performance

In each round of this selection, a candidate that most worsened the performances was eliminated from the feature set.

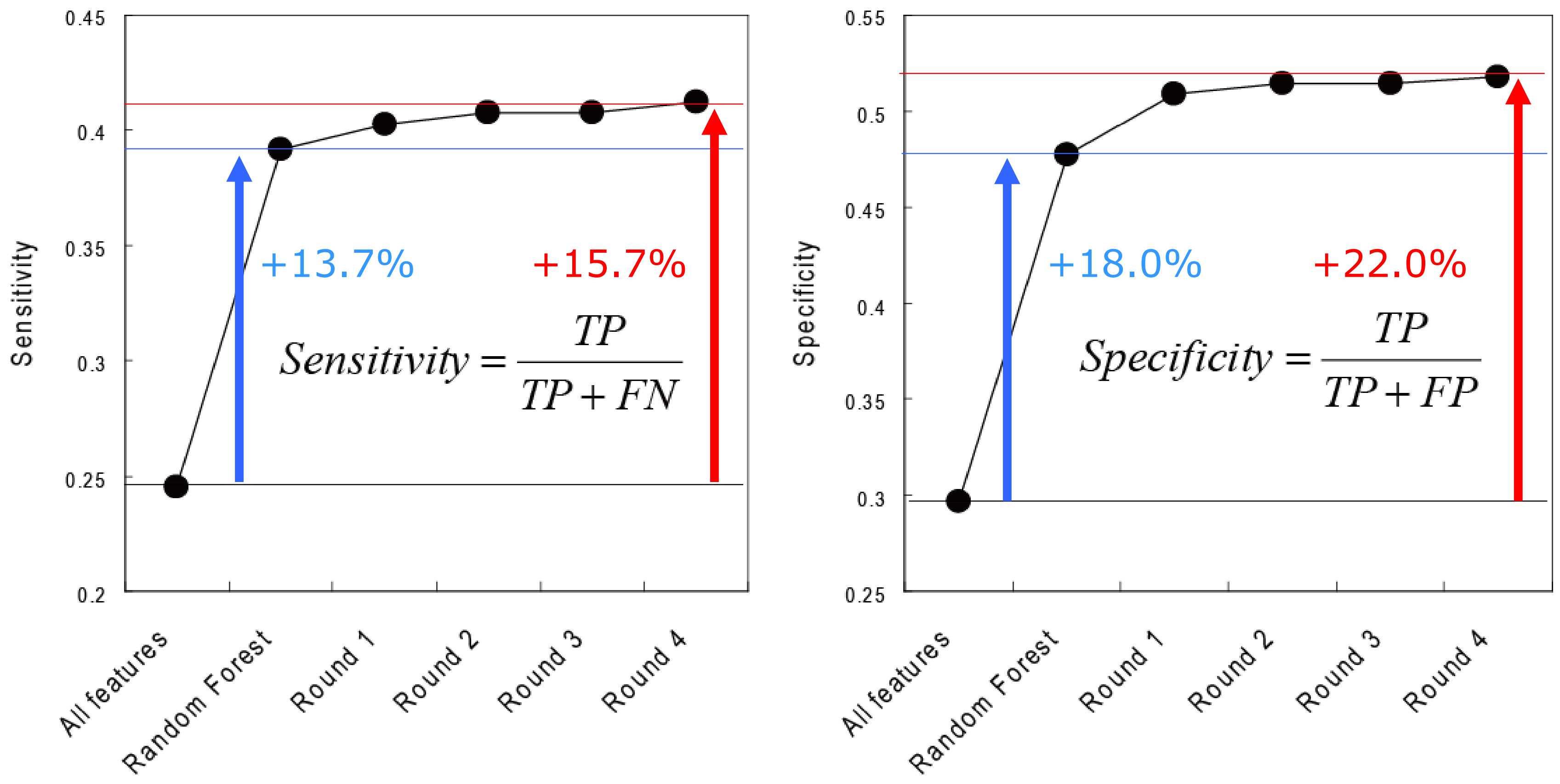
Results

Importance score of the feature candidates





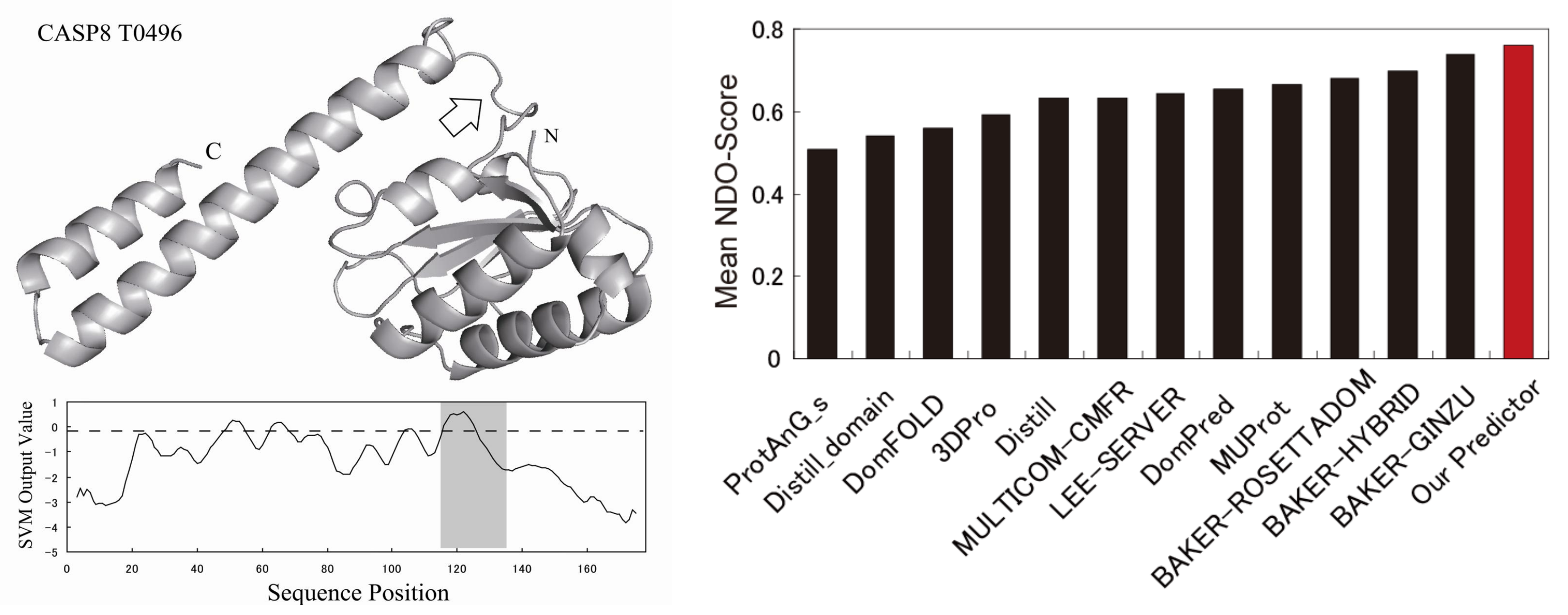
Improvement by feature selections



$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TP}{TP + FP}$$

Compare with CASP8 servers



Computational Time of the Feature Selection

	Runing Time (hour)	Feature Total	hours/Feature
Random Forest	20	2870	0.007
Backward Selection	100	47	2.128

Conclusion

- The combination of random forest & backward selection efficiently determined the optimal features.
- The prediction performances of our predictor improved by over 15% by the feature selection.