RUHR UNIVERSITÄT BOCHUM

RUB

A Versatile Combination of Classifiers for Protein Function Prediction

Haneen Altartouri
(haneen.altartouri@ini.rub.de)

Tobias Glasmachers

Institute for Neural Computation
Ruhr University Bochum

27/9/2020

About the presenter

❑ Haneen Altartouri

- PhD student in Institute for Neural Computation, Ruhr University Bochum.
- Member of "Theory of Machine Learning" group lead by Tobias Glasmachers.
- My research is focused on improving the prediction of protein functions.

Contents

Introduction

- Motivation of protein classification

High-Throughput technology

discover

Millions of sequences

EQEDVMIEGVEKFFS
EQVDLGPVLNVLKWH
ETVEKIVDQYREPVK
EVLKLKDEVRLSTRE
EVSLHEVKDFDPDVL
EYAERQGKTPITLVD
FASEFKSRFFIWSQE
AAEWVLAYMLFTKFF
ACMLDGGNMLETIKV
ACVKDLVSKYLADNE
AEHDRQVLNNLSNCV
AERTVTVRRVGPGGRAV
.....

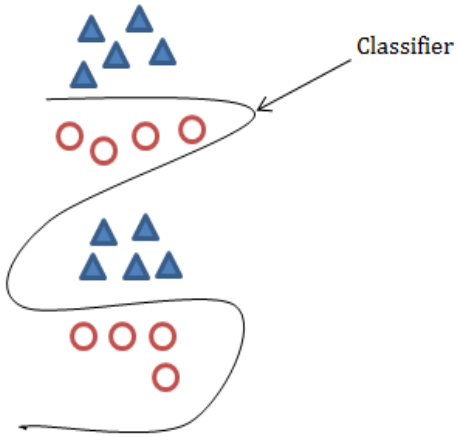
Sequences with unknown functions !!!

Computational approaches to solve various protein prediction problems in a faster and more cost-effective manner.

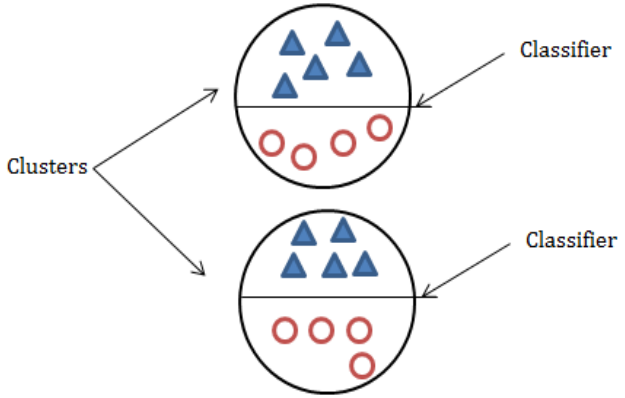
Introduction (cont.)

❑ Clustering before the classification

- Classification is often easy if the discriminative features are homogeneous for the whole data set.
- For heterogeneous datasets, we should therefore find homogeneous regions and address them with separate classifiers.

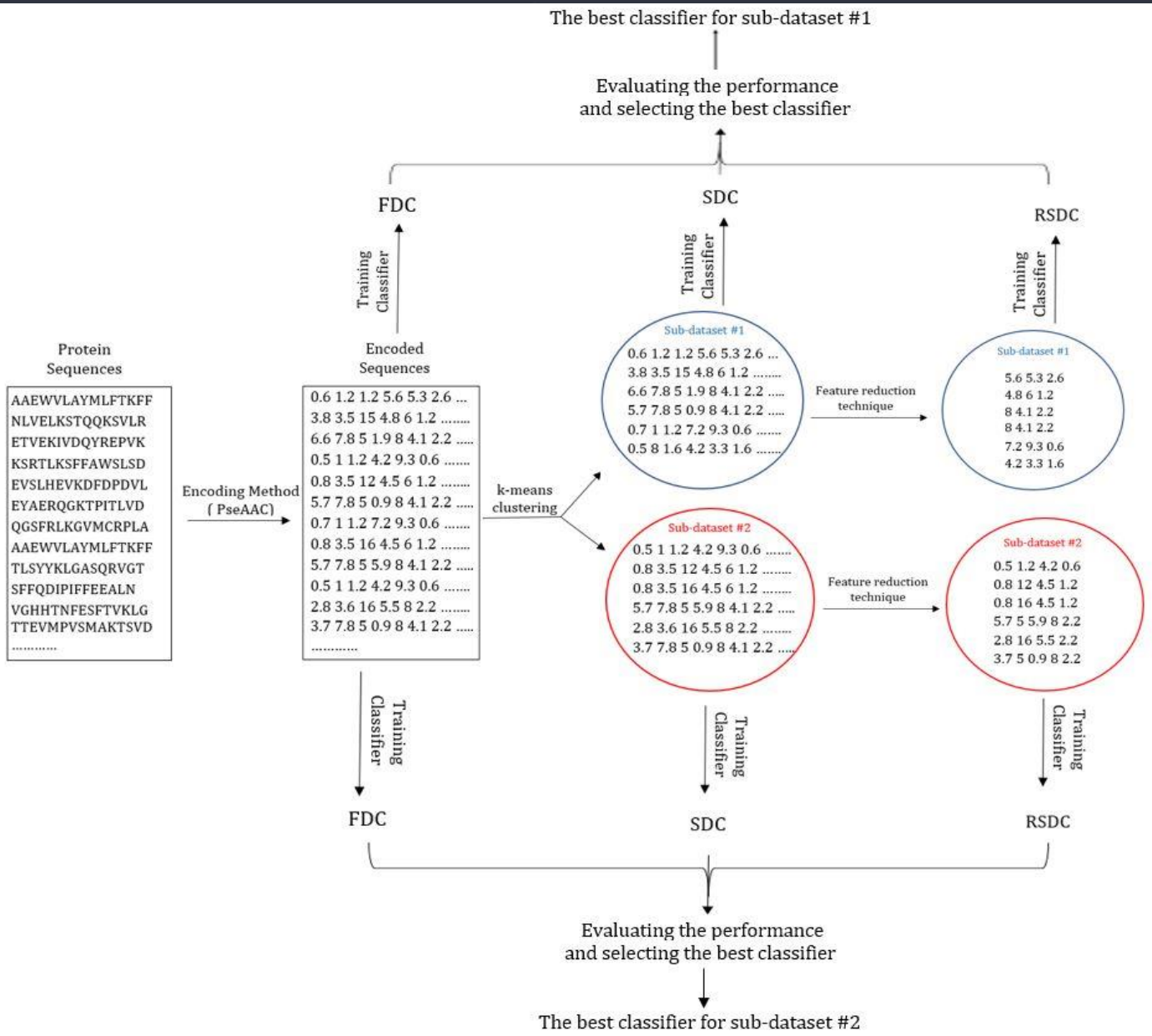


Classification without clustering



Classification based on clustering

The proposed approach



The proposed approach (cont.)

❑ Representing Protein Sequences

- We used Chou's Pseudo Amino acid Composition (PseAAC) descriptors.
- Two sets of Physico-chemical properties (PCPs) were tested:
  1. 3 PCPs (hydrophobicity, hydrophilicity, and side chain mass).
  2. 50 non-redundant PCPs of amino acids.

❑ Clustering dataset into sub-datasets

- K-means was used.
- We tuned the number of sub-datasets (k) for each dataset, to study its effect on the proposed approach.

The proposed approach (cont.)

❑ Reducing Feature Vector Dimensionality

- Two reduction techniques were tested:
  1. Recursive Feature Elimination (RFE).
  2. Principal Component Analysis (PCA).

❑ Classifier Selection

- For each sub-dataset we have up to three classifiers available: FDC, SDC, and RSDC.
- We estimate the performance of all three classifiers by means of cross-validation.
- We select the classifier with highest AUC.

Benchmarks

Dataset	# of Positives	# of Negatives
DNA-binding proteins	523 binding proteins	543 non binding proteins
Antioxidant proteins	250 antioxidant	1547 non-antioxidant
RNA-binding proteins	2780 binding proteins	7077 non binding proteins
Antimicrobial peptides (AMP)	869 AMPs	2405 non-AMPs
Caspase 3 human substrates	247 cleaved peptides	247 non-cleaved peptides
Major Histocompa. Complex II (MHCII)	3510 binding peptides	1656 non-binding peptides

Experiments and Results

❑ Selecting the best Classifier

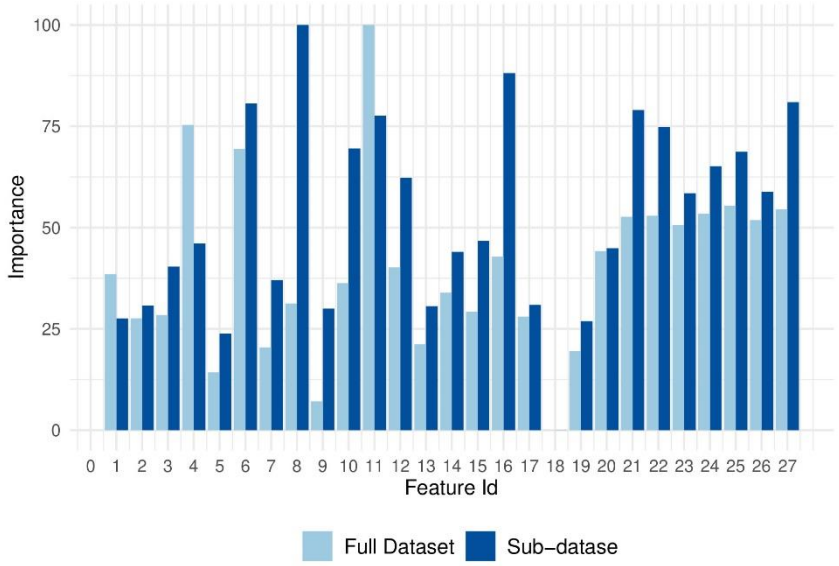
- Different classifiers were tested on the full datasets (FDCs): SVM, RF, ANN, and xGBoost.

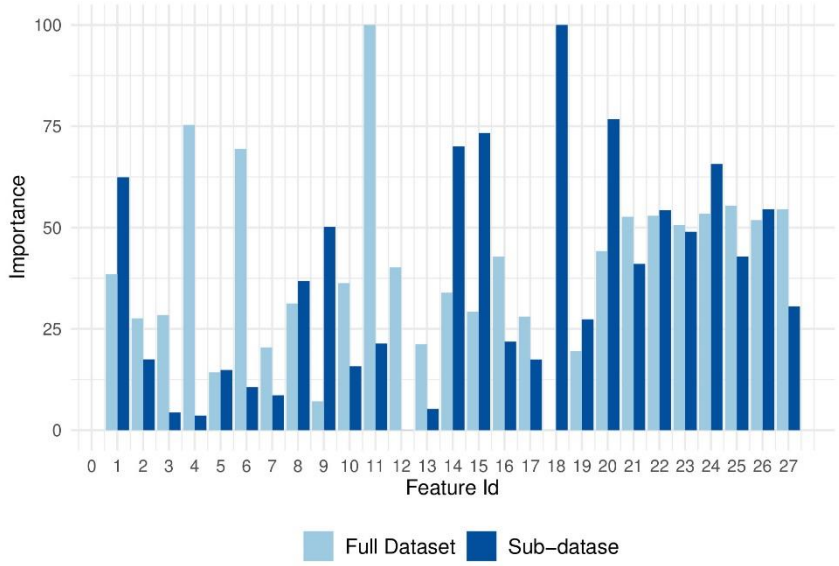
- The results showed that:
  1. SVM is the best choice for most datasets using 50 PCPs.
  2. RF is the best choice when using 3 PCPs.

Experiments and Results (cont.)

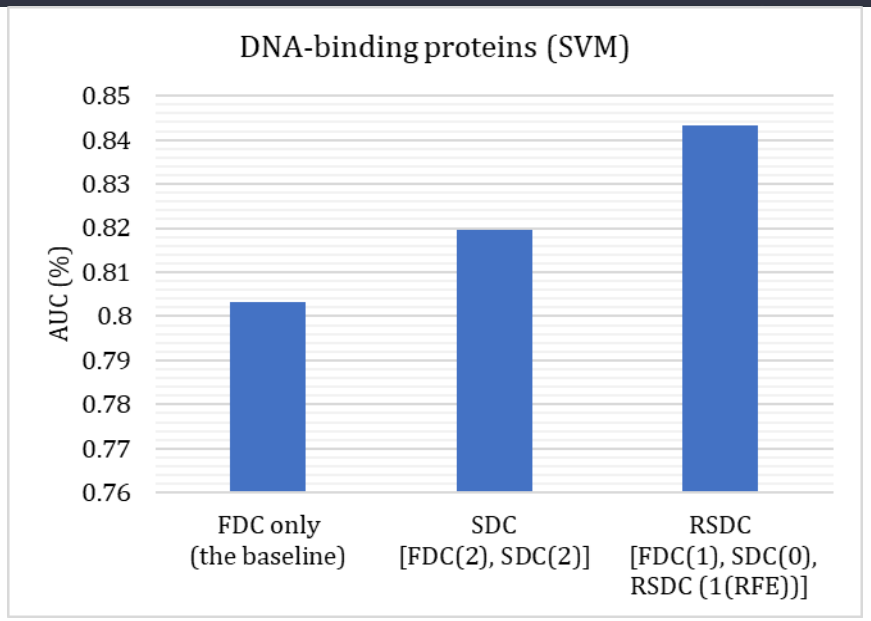
❑ Applying feature reduction on sub-datasets

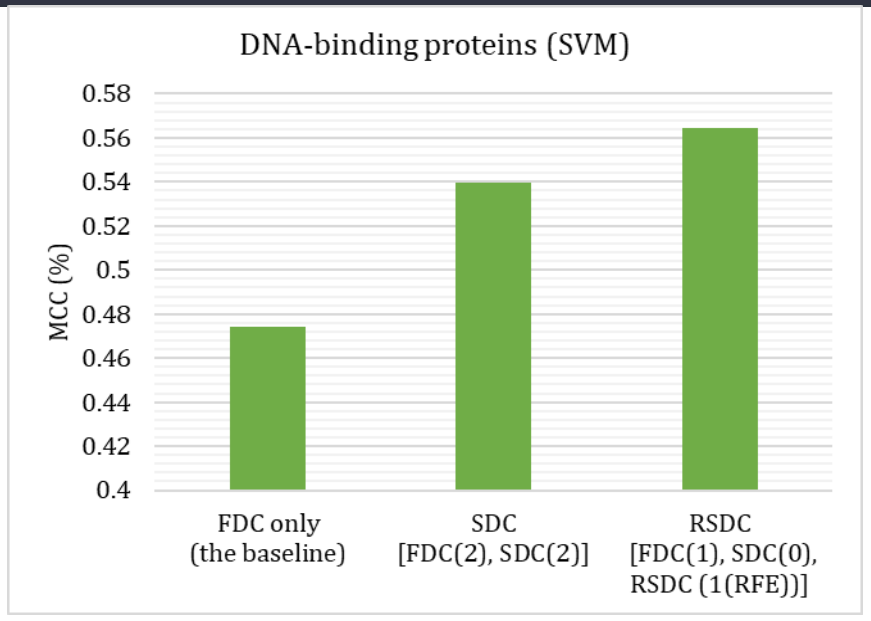
- The importance of the features differs not only between the two sub-datasets, but also from the full dataset.
- Therefore, applying feature reduction on a per-cluster basis has the potential to improve overall performance.

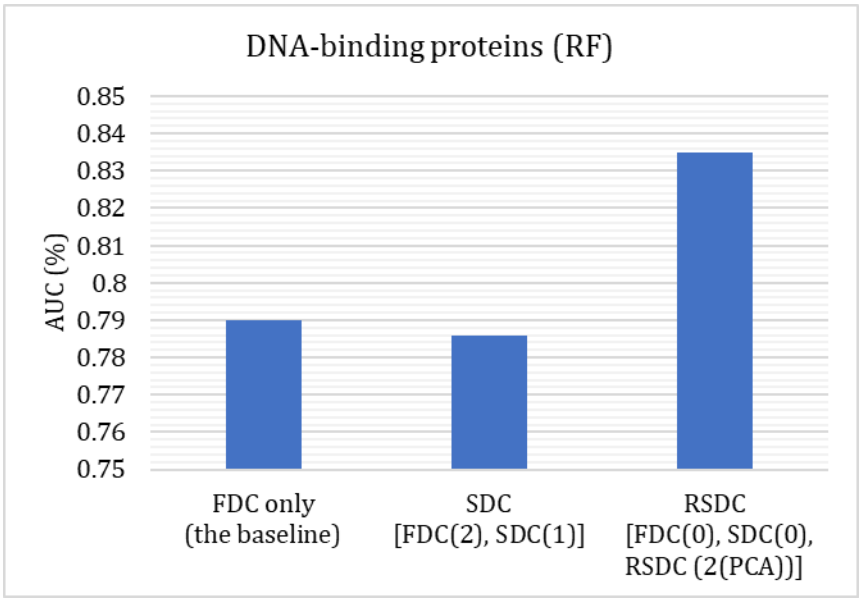


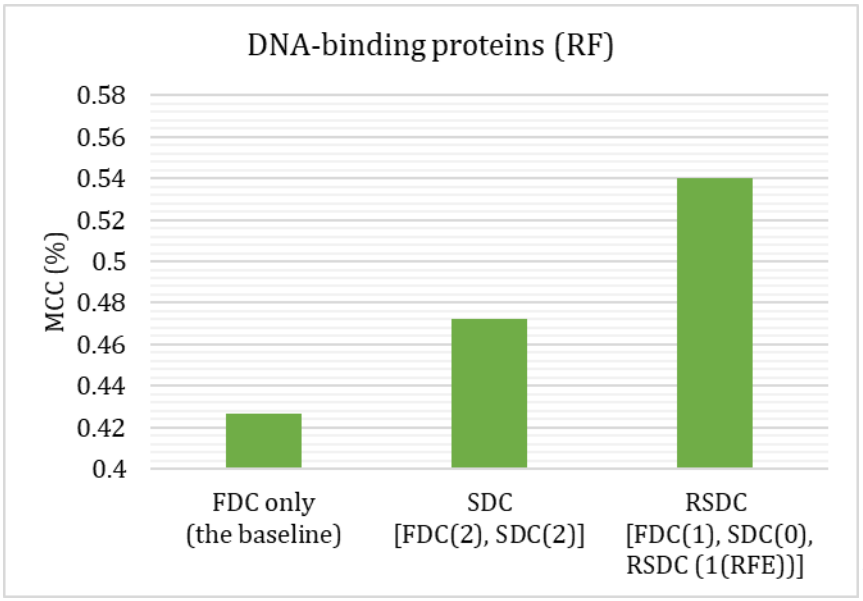


Experiments and Results (cont.)

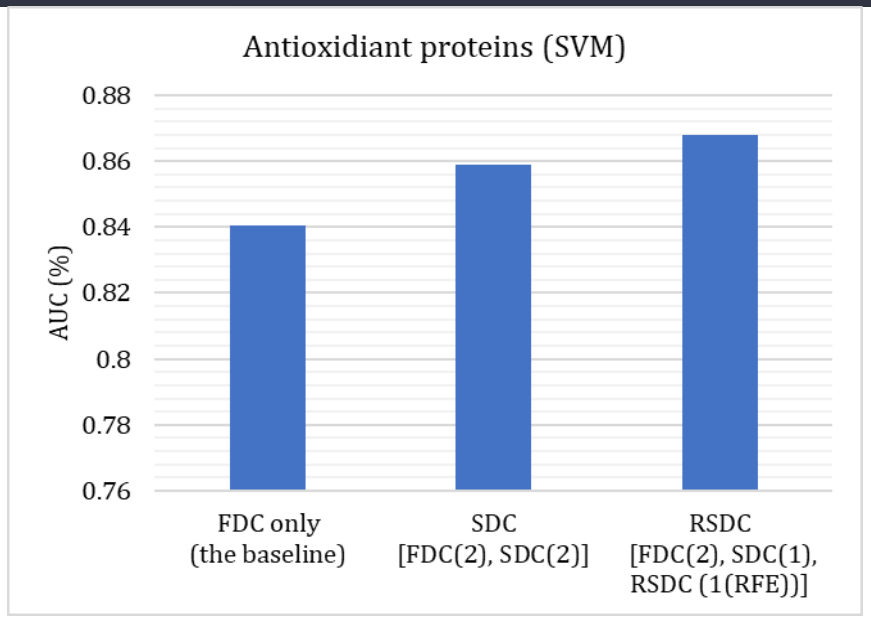


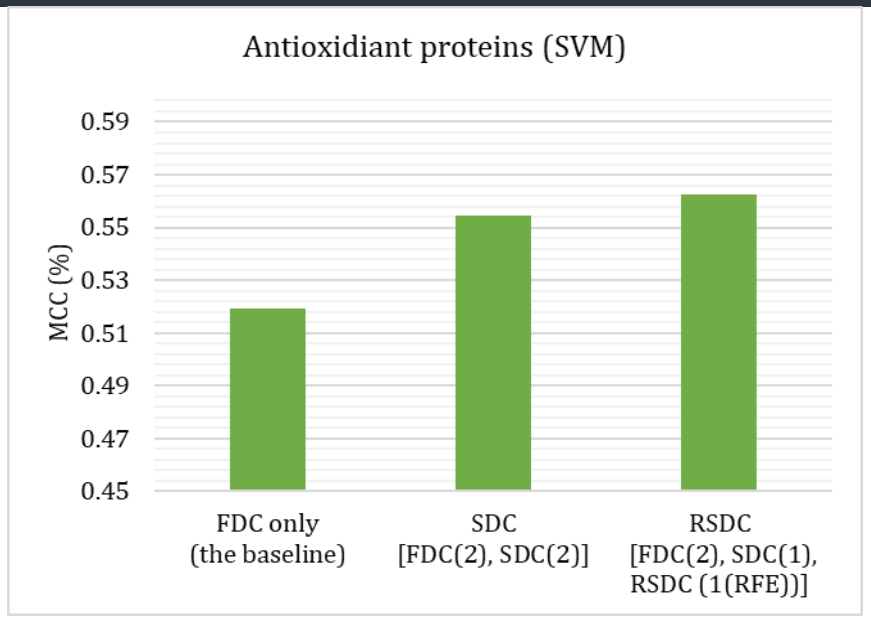


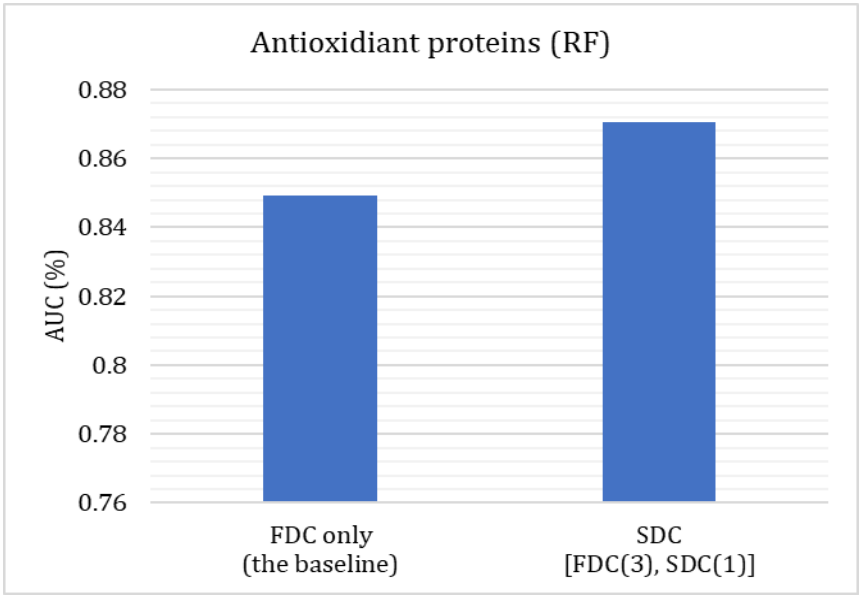


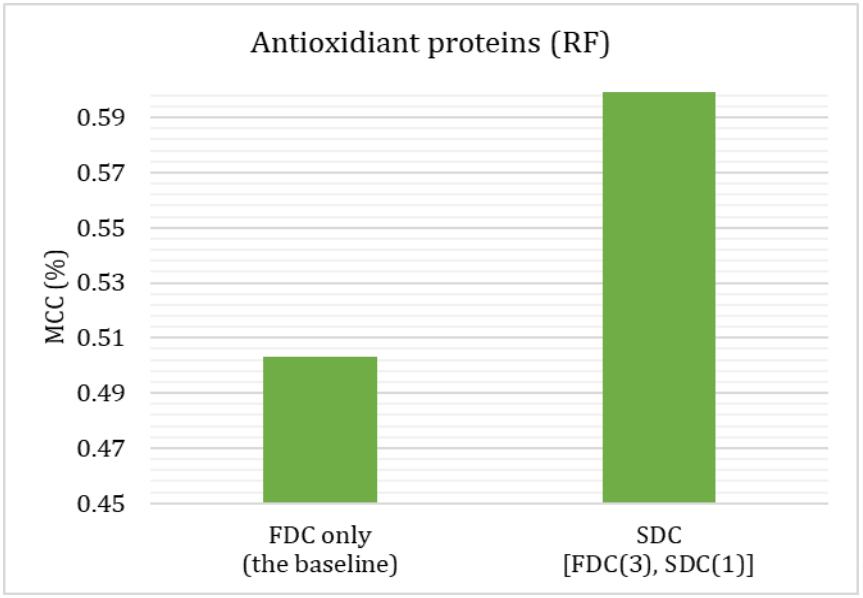


Experiments and Results (cont.)

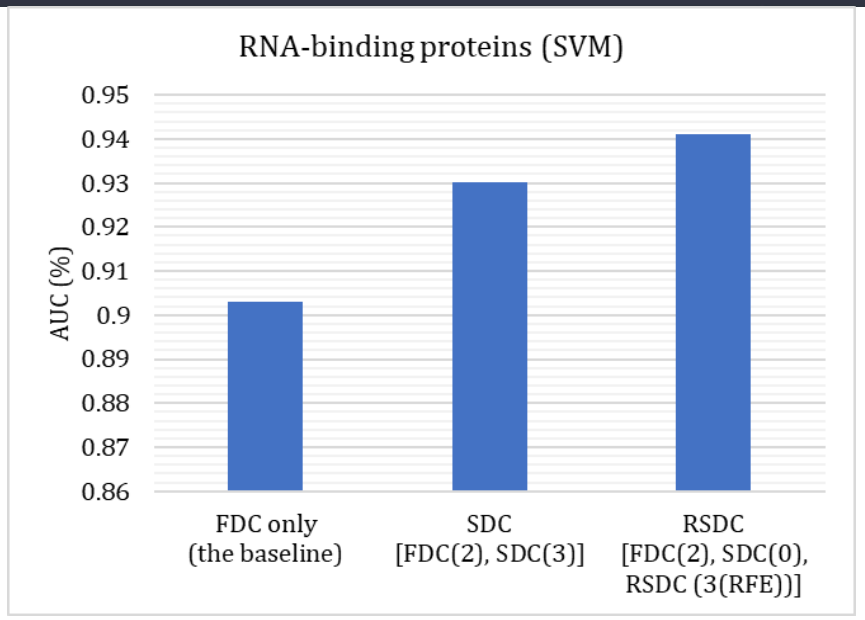


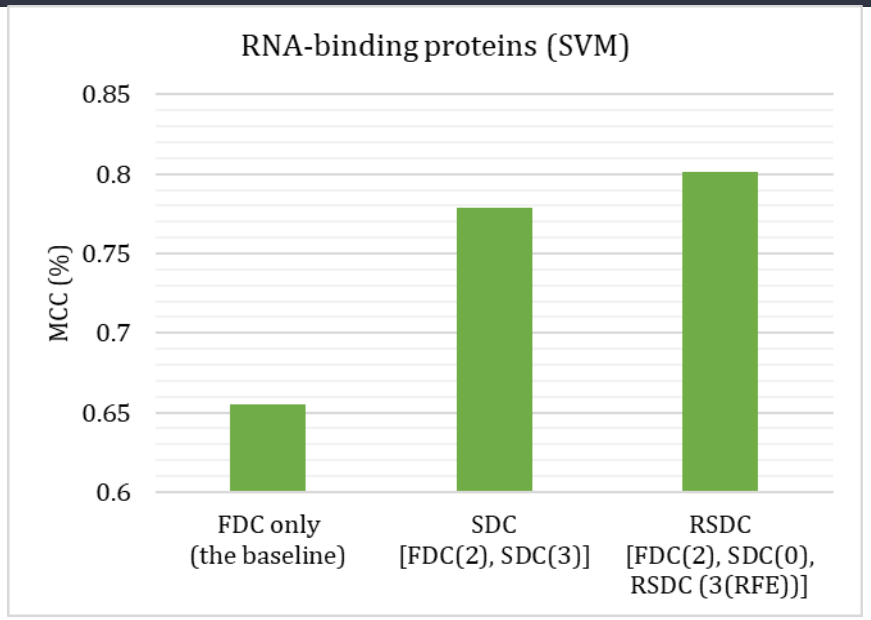


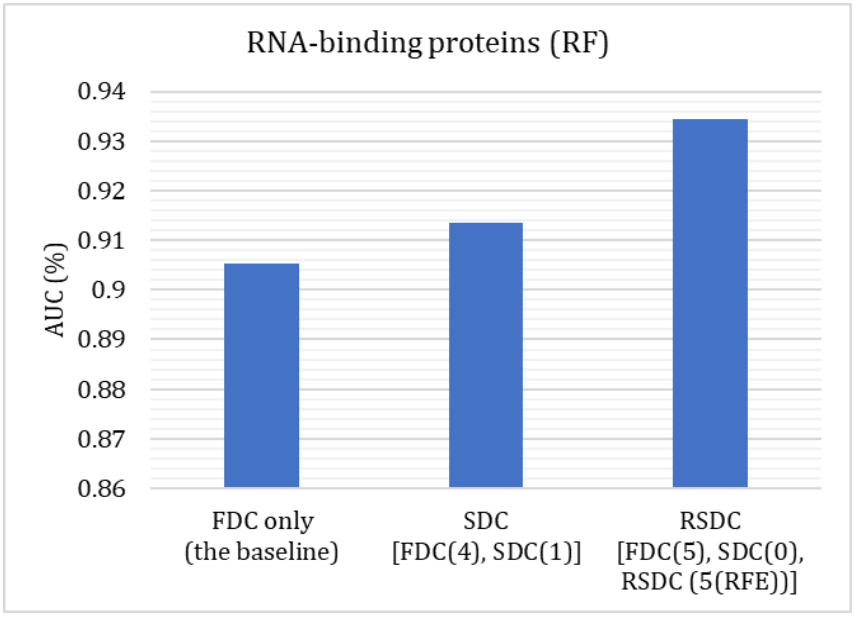


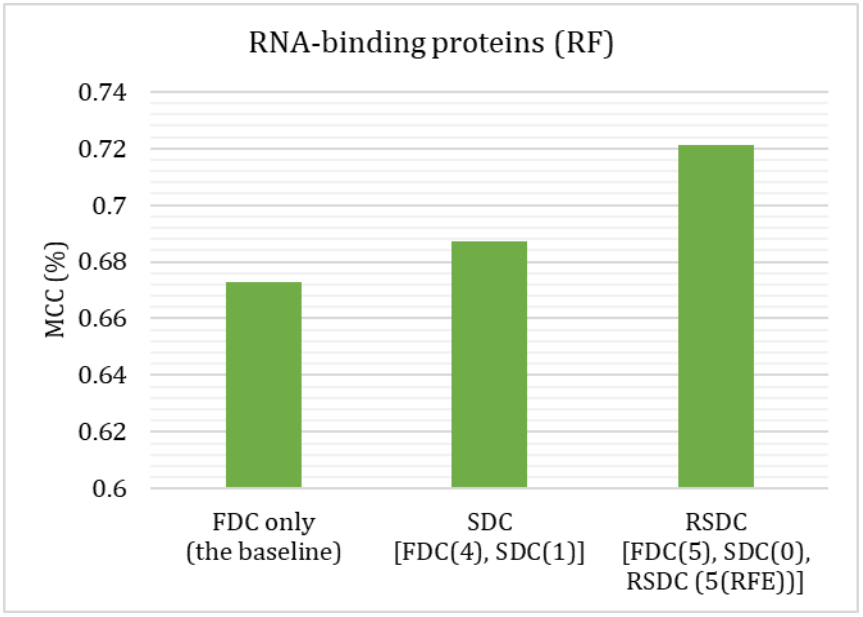


Experiments and Results (cont.)

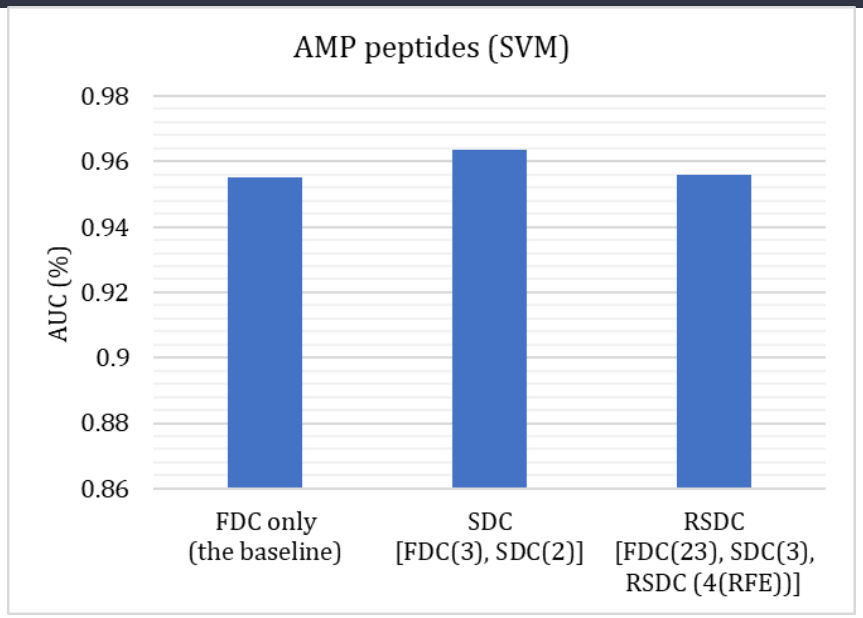


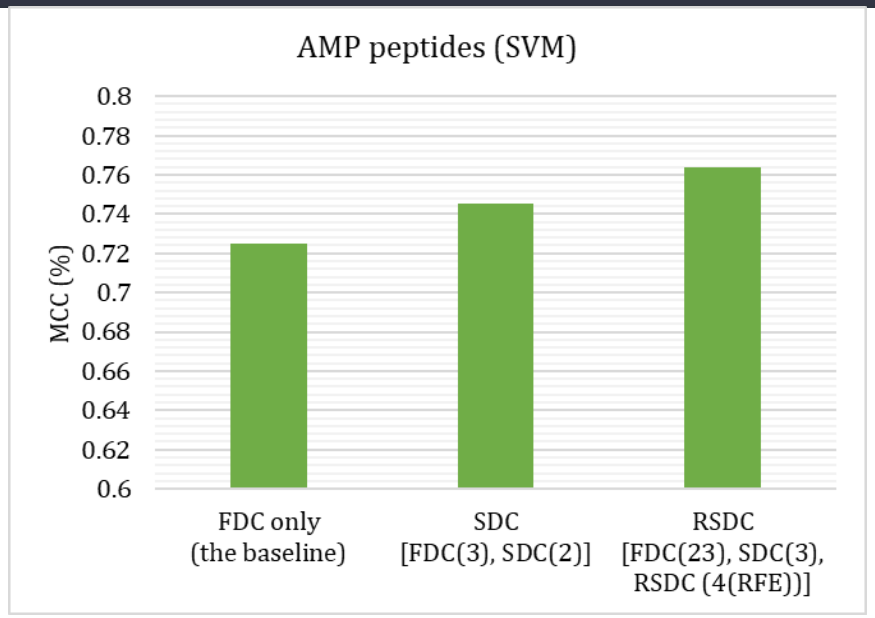


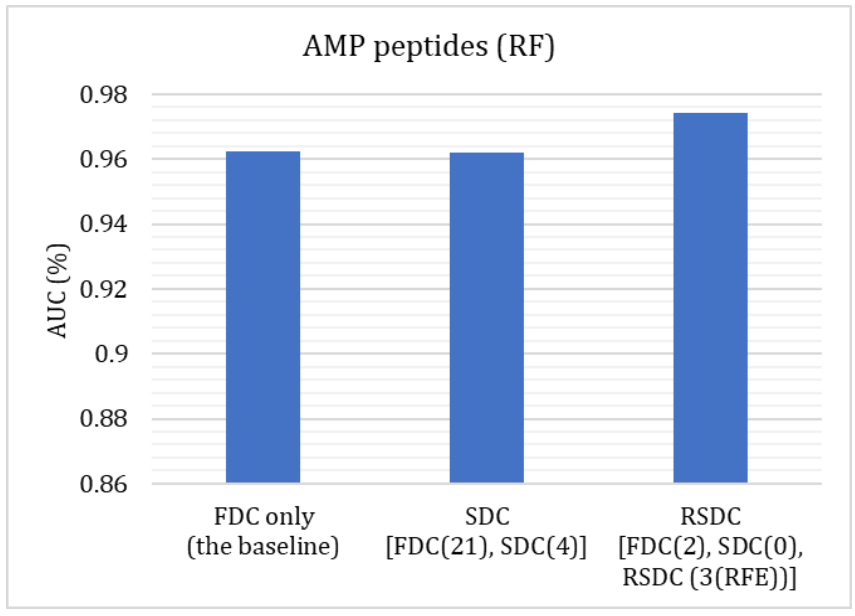


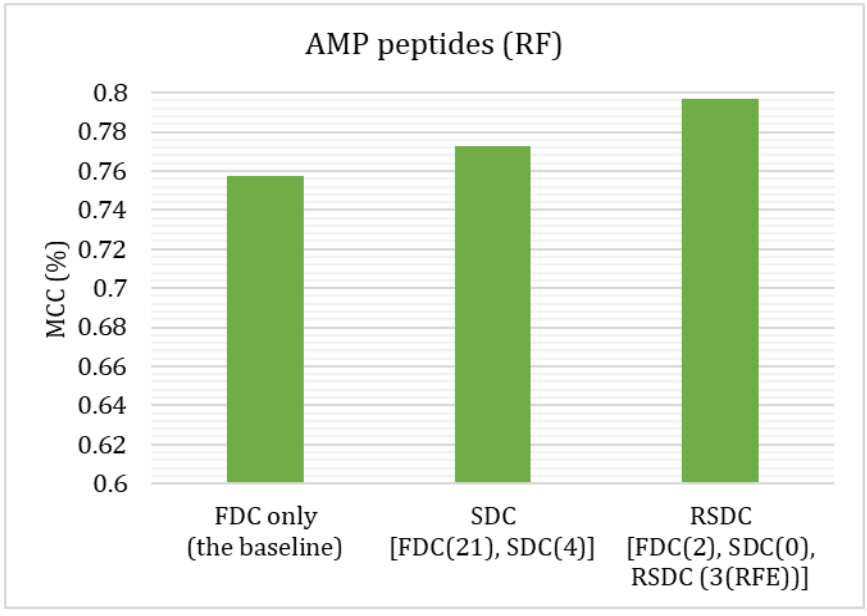


Experiments and Results (cont.)

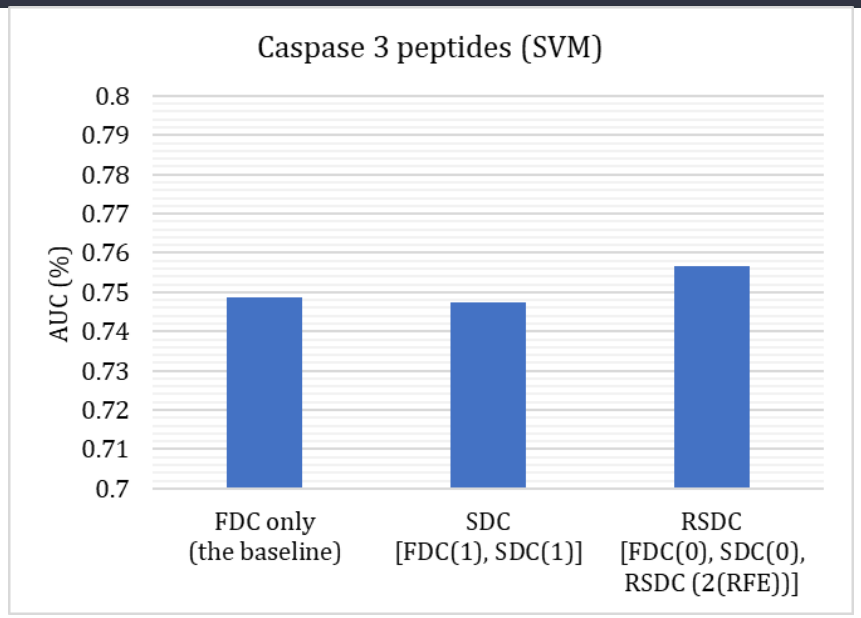


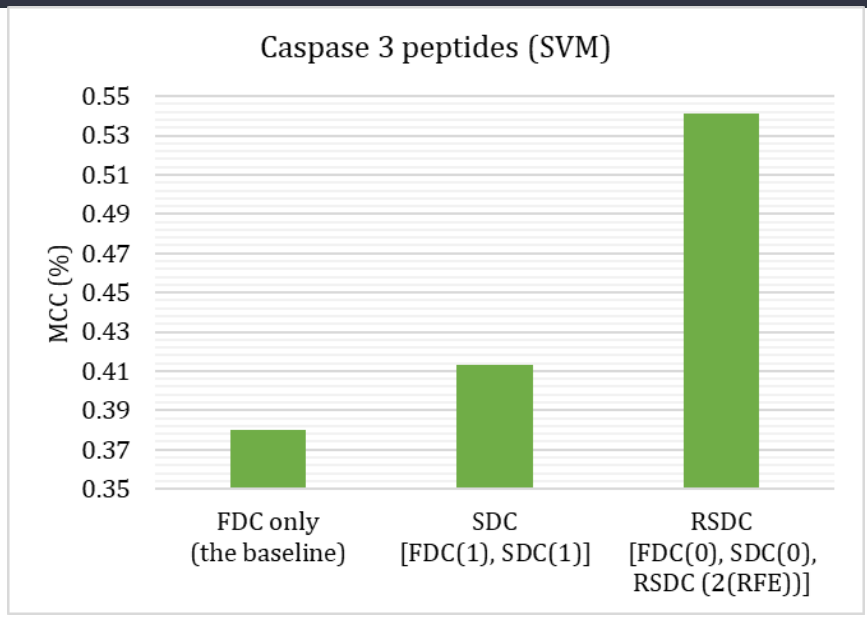


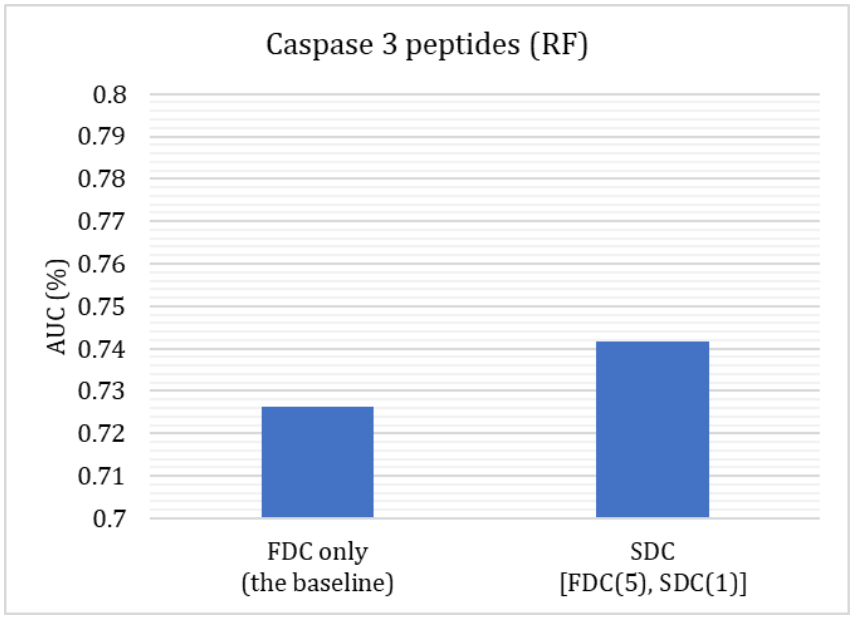


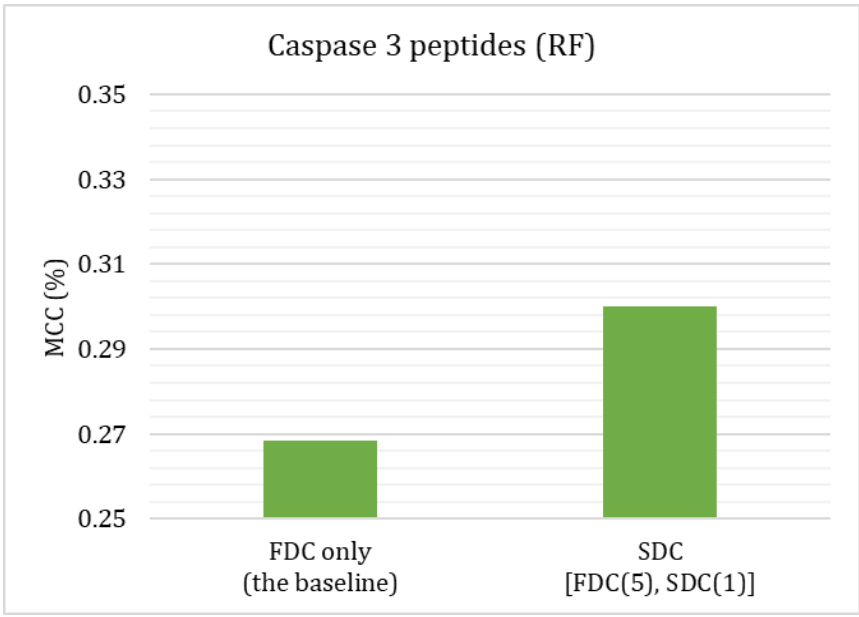


Experiments and Results (cont.)

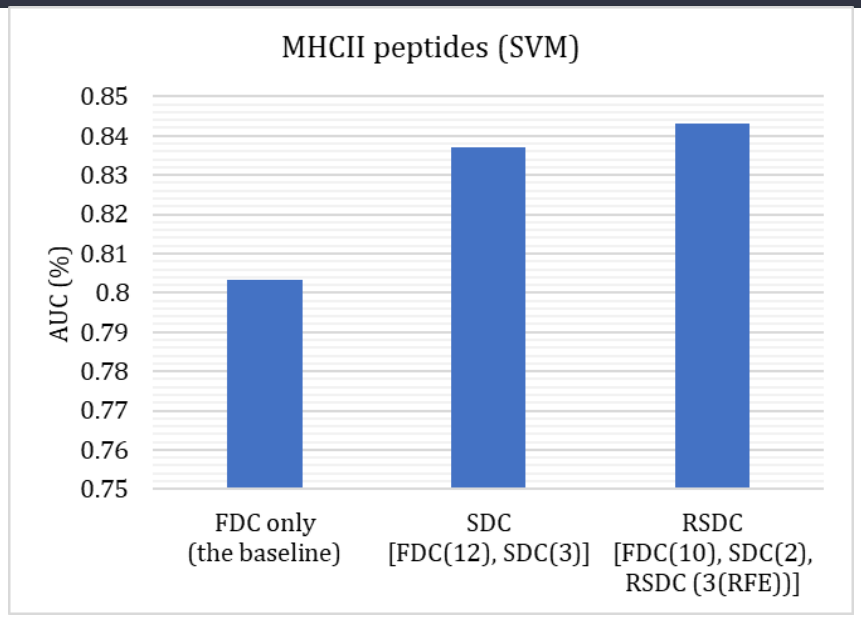


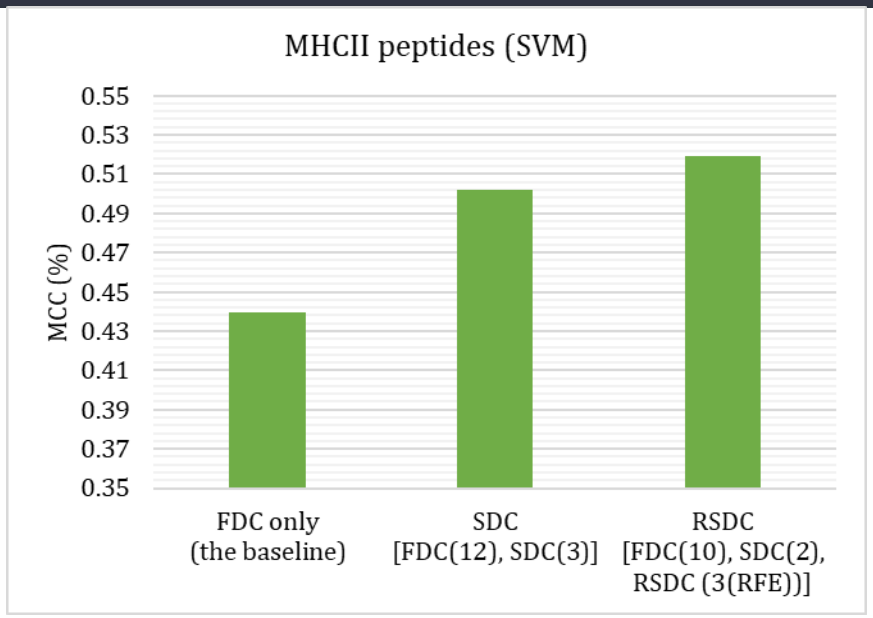


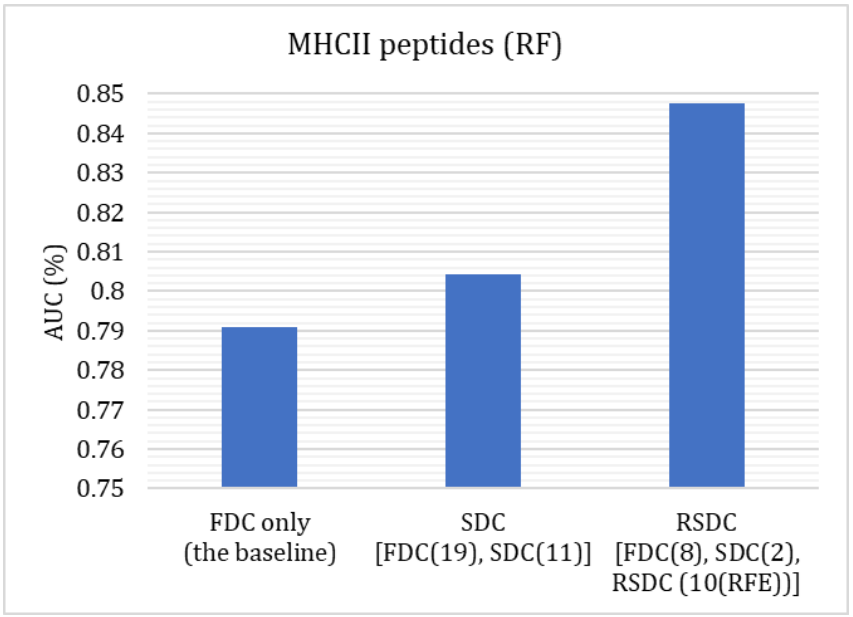


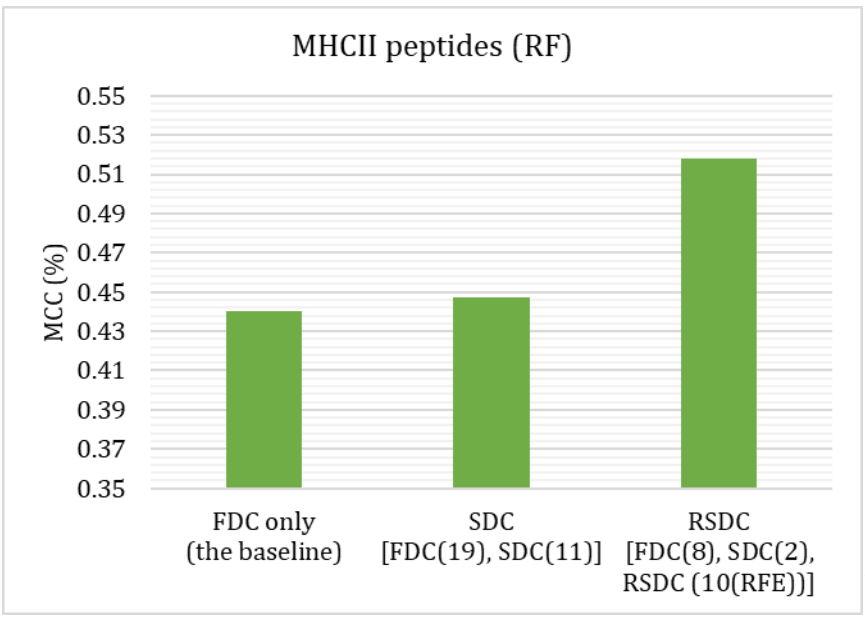


Experiments and Results (cont.)



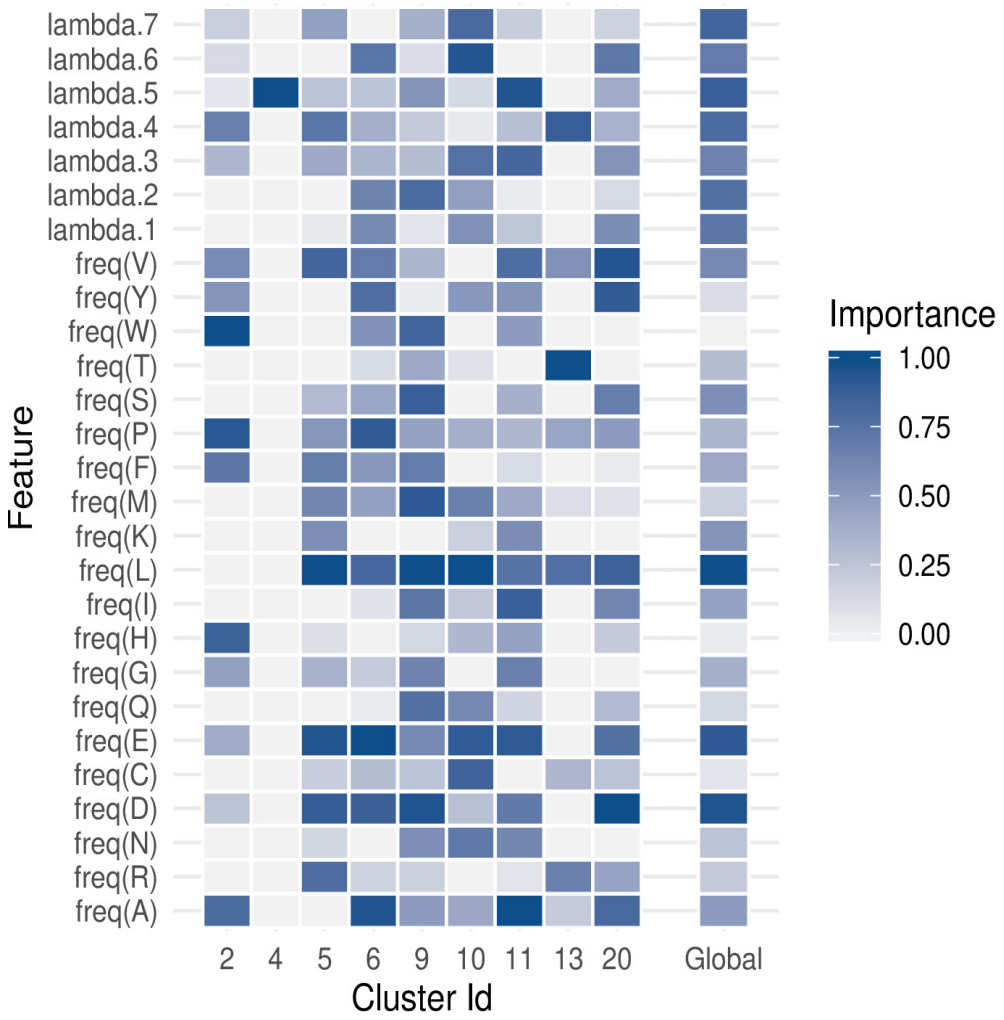






Experiments and Results (cont.)

- In most cases, RFE shows that the frequencies of amino acids play an important role in classifying the sequences inside the clusters, while the sequence order has a higher impact on classifying the full dataset.



Experiments and Results (cont.)

- For datasets containing long protein sequences, RFE shows that the optimal sets of features for clusters contain only a bit more than 50% of all available descriptors.

Conclusion

- We have studied the effect of exploiting homogeneous sub-datasets inside protein sequence data by training multiple classifiers on sub-datasets.

- The proposed approach handles each sub-dataset as a separate classification problem that requires tuning the hyper-parameters and finding the best features separately.

- We have evaluated the performance of SVM and RF classifiers inside the sub-datasets, and RFE and PCA are tested as a reduction feature algorithms.

- SVM and SVM-RFE achieved good performance for most datasets.

Conclusion (cont.)

- The performance of the proposed approach depends on the number of sub-datasets, the encoding method, and for each cluster the classifier with its hyperparameters and the feature reduction method applied.

- The results indicate that the proposed approach improved the overall performance of function prediction of protein sequences in the most cases.

- Results indicate that many protein sequence datasets suffer from heterogeneity.

RUHR
UNIVERSITÄT
BOCHUM

RUB

Thank you