
DTU

Paul Jeremy Simon (s202592), Maria Vinh Thuy Tien Ta (s134042), Huijiao Yang (s202360), Group 11

Comparative study of PSSM, SMM and ANN

Outline

1. Introduction
2. Materials and methods
3. Results
4. Discussion

Introduction

What does exist already?

- High performance prediction tools for MHC class I binding

Challenges with this approach

- Low explanation and understanding of expectancy

Our project

- Use nested 5-fold cross-validation to evaluate the performance of three prediction model: PSSM, SMM and ANN
- Which method bring the best performance
- Investigate strengths and weaknesses of each models
- Robustness test (distribution and evaluated metric)

Materials and methods (dataflow)

```
WLSLLVPFV 0.212813 False 1.4863802827668746 -0.04297123071848948 0.10720820690692744
YPAEITLTW 0.084687 False 0.6471137919867596 0.12384865530977143 0.0758516468520803
GRKTPLLCF 0.084687 False 1.4898622433715834 -0.15178250087766265 0.06859840360137442
AQQFCQYLI 0.040183 False -1.5901644659920737 -0.04651559693082542 0.07423037934453822
RSARASSRY 0.405446 False 6.770392392040943 0.2952533254924411 0.32967375054015
```

Train/Test (SMM and ANN)
Matrix construction (PSSM)

Evaluation

Metrics calculation

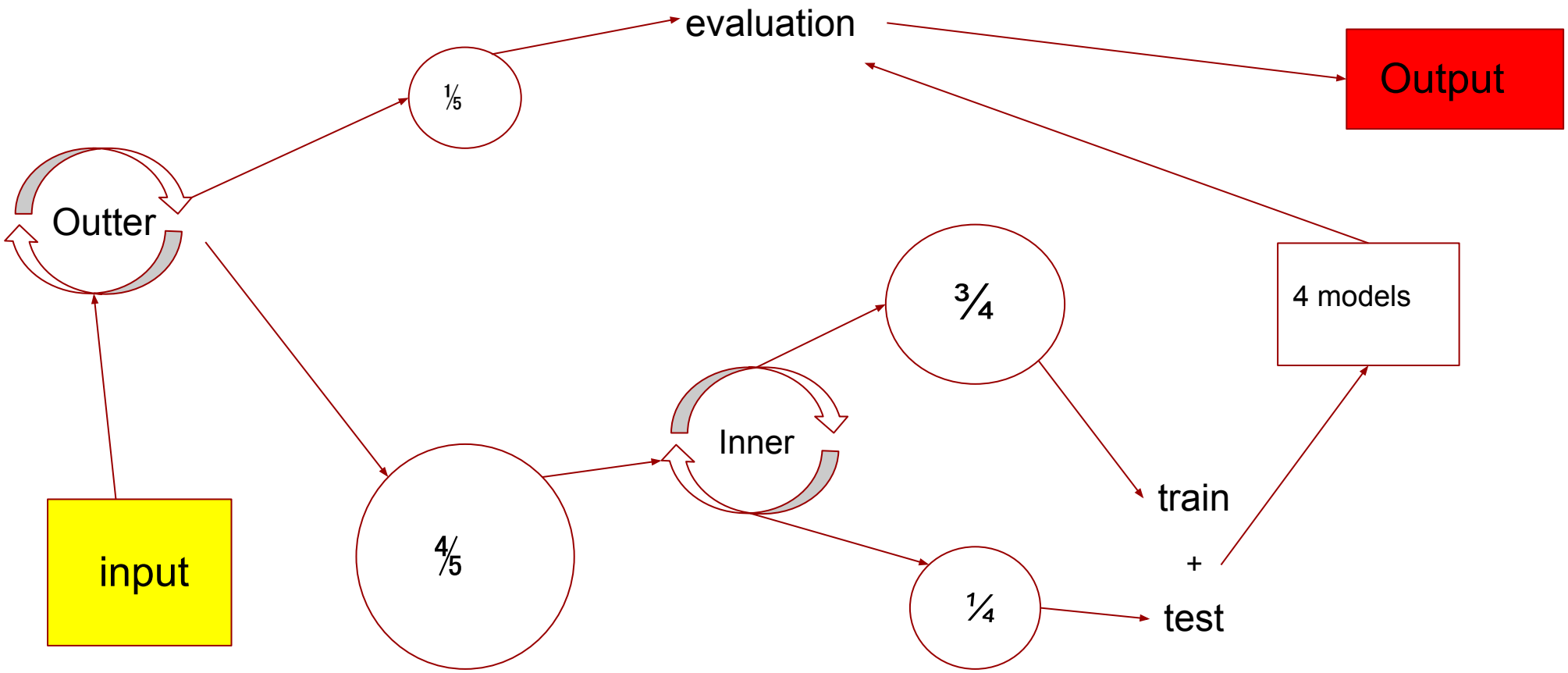
Raw data → Models → Predictions → Performance (check by distribution)

```
ATDALMTGY 1.000000 A0101
CTDDNALAY 0.860988 A0101
LTDDMIAAY 0.820152 A0101
CTELKLSDY 0.770337 A0101
STDHIPILY 0.744393 A0101
```

- PSSM Matrix
- SMM weights
- ANN weights

A3301	184	4.53283155283916	0.27895216579629206	0.16123577045586732
A0201	1181	4.742624881768609	0.35437928934027096	0.15114507329884716
A0203	639	4.384372926268005	0.3567661514491745	0.16790310313668377
A0206	514	4.280222931684366	0.3212191517507646	0.1667178969281103
A0301	517	4.448145900433608	0.24886532638853684	0.1478474496560525

Materials and methods (cross-validation)

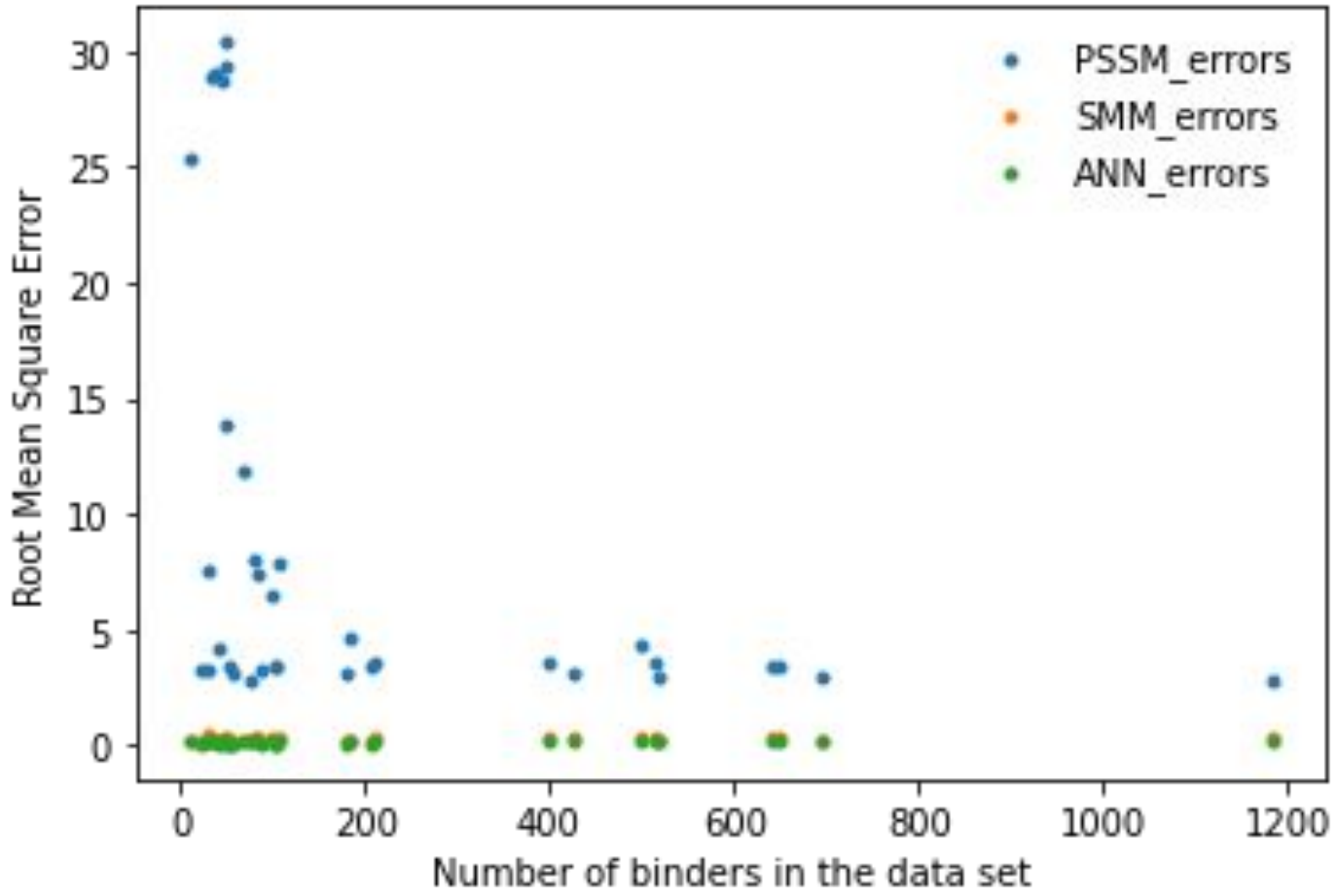


Results

RMSE plot

PSSM: 3.93 - 29.78
SMM: 0.17 - 0.47
ANN: 0.08 - 0.24

ANN > SMM > PSSM **35** times



Results

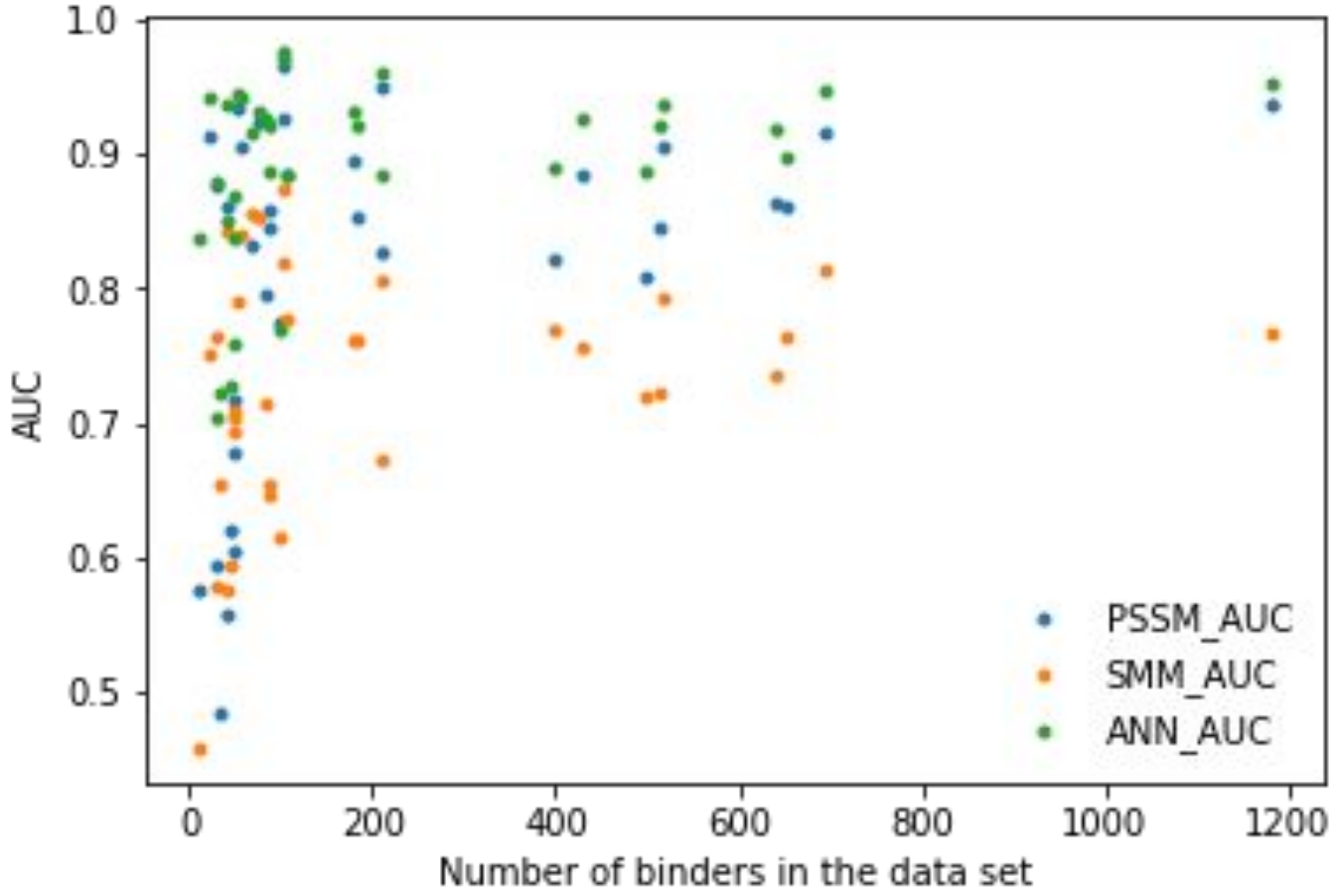
AUC plot

PSSM: 0.48 - 0.97
SMM: 0.46 - 0.87
ANN: 0.70 - 0.98

ANN > SMM > PSSM **5** times

ANN > PSSM > SMM **28** times

PSSM > ANN > SMM **2** times



Results

Accuracy plot

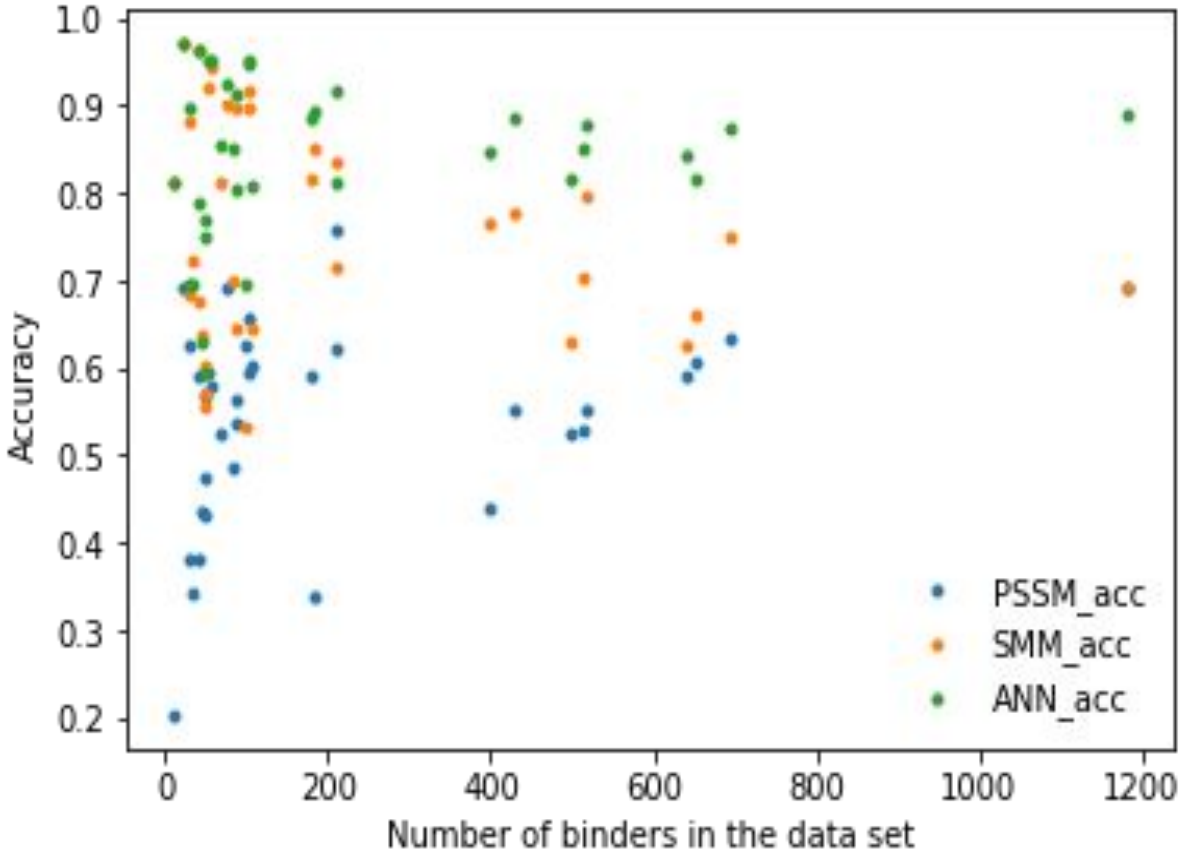
PSSM: 20.3% - 75.9%
SMM: 53.3% - 97.0%
ANN: 59.6% - 97.0%

ANN > SMM > PSSM **28** times

ANN > PSSM > SMM **3** times

SMM > ANN > PSSM **2** times

PSSM > SMM > ANN **2** times



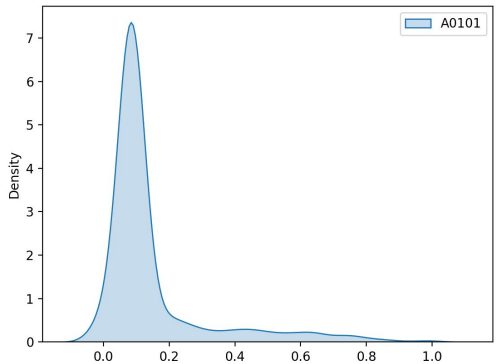
Results

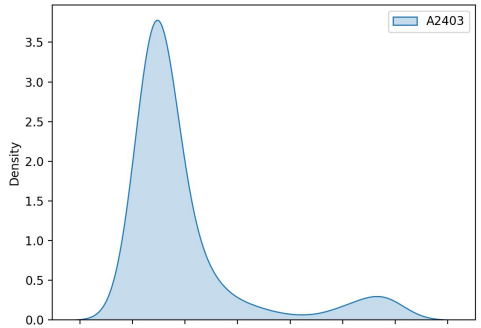
Distribution types:

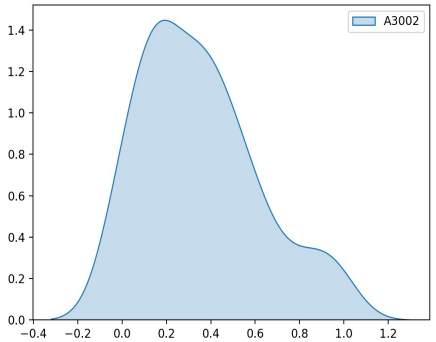
Nest: Single Peak or Higher Peak point above **3.0**

Width: Double Peak (difference not significant) or Wide shape (below **3.0**)

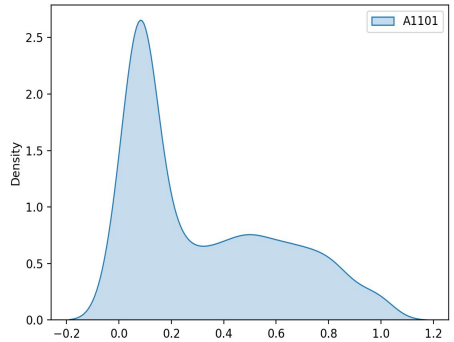
Nest

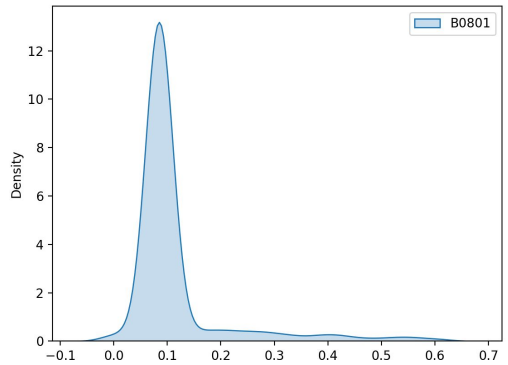


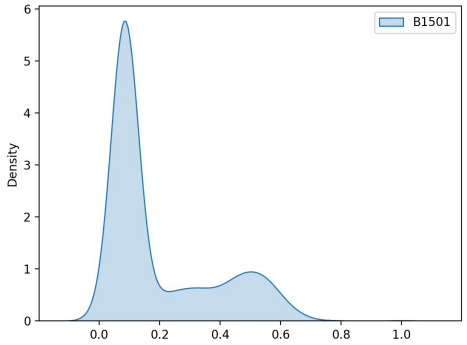


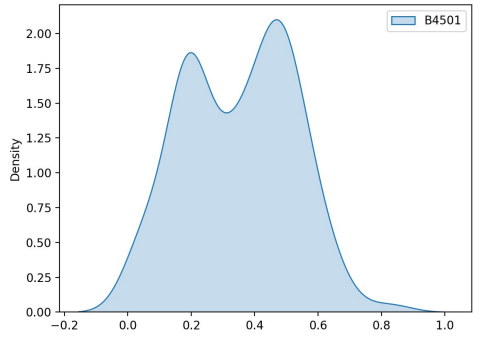


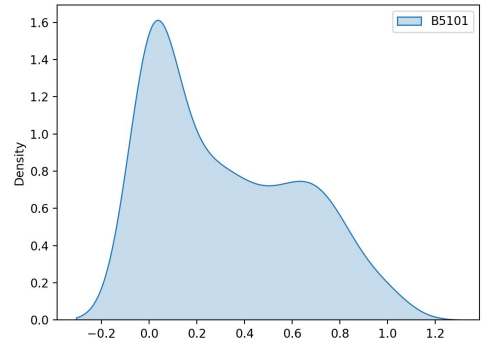
Width









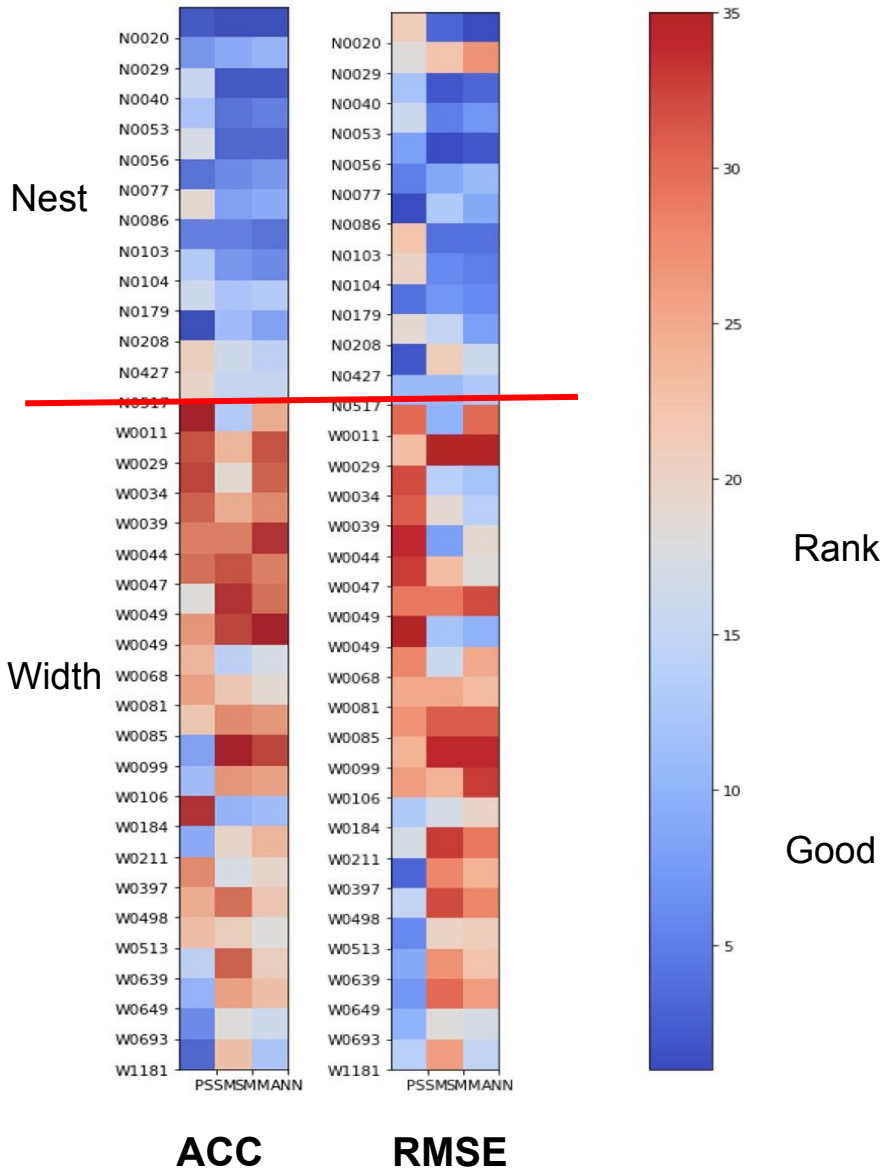


Result

Similarity between the 2 metrics for top 10:
ANN: 80%; SMM: 60%; PSSM:30%

Top 10 rank table

	Accuracy				RMSE	
PSSM	SMM	ANN	Rank	PSSM	SMM	ANN
B0702	B0801	B0801	1	A6901	B2705	B0801
B0801	B4001	B4001	2	A3101	B4001	B2705
A0201	B2705	B2705	3	A6802	B0801	B4001
A3001	A2601	A0101	4	B1501	A0101	A0101
A0101	A0101	A2601	5	A3001	A2601	B5801
A1101	A3001	B5801	6	A0206	B5801	B1501
A2403	B5801	A3001	7	A0202	B1501	A2601
A2402	A6901	B0702	8	B2705	B4402	B0702
B3501	A2403	A6901	9	A0203	A3001	A6901
A0202	A3301	A2403	10	A1101	B5701	B4501
50 %	**90 %**	**100 %**		50 %	**80 %**	**90 %**



Discussion

No hyperparameters tuning during the project

- Machine learning methods might have better performance but SMM didn’t

Each metric focuses on a precise performance

- RMSE is a regression metric, to measure
- Accuracy/AUC are focusing on classification

Robustness of Distribution impacts and Evaluated methods

- All is nest preferred but PSSM relatively stable in different distribution
- ANN and SMM are stable in different evaluated methods, but PSSM is not in the same file

SUMMARY

- ANN has the best in performance and robustness in different evaluation
- ANN and SMM are sensitive to distribution
- PSSM is bad in performance and robustness but not sensitive to distribution

Other focuses can be investigated

- How well do the methods predict which position in the peptide is the most important?

References

- Bjoern Peters, Huynh-Hoa Bui, Sune Frankild, Morten Nielsen, Claus Lundegaard, Emrah Kostem, Derek Basch, Kasper Lamberth, Mikkel Harndahl, Ward Fleri, Stephen S Wilson, John Sidney, Ole Lund, Soren Buus, and Alessandro Sette. A community resource benchmarking predictions of peptide binding to mhc-i molecules.PLOS Computational Biology, 2(6):1–11, 06 2006.
- Marek Wieczorek, Esam T. Abualrous, Jana Sticht, Miguel Alvaro Benito, Sebastian Stolzenberg, Frank Noe, and Christian Freund. Major histocompatibility complex (mhc) class i and mhc class ii proteins: Conformational plasticity in antigen presentation. Frontiers in Immunology, 8:292, 2017.
- Yohan Kim, J. Sidney, S. Buus, A. Sette, M. Nielsen, and Bjoern Peters. Dataset size and composition impact the reliability of performance benchmarks for peptide-mhc binding predictions. BMC Bioinformatics, 15, 2014.
- Limin Jiang, Hui Yu, Jiawei Li, Jijun Tang, Yan Guo, and Fei Guo. Predicting MHC class I binder: existing approaches and a novel recurrent neural network solution. Briefings in Bioinformatics, 06 2021.
- Morten Nielsen. Stabilized matrix method. http://www.cbs.dtu.dk/courses/27625.algo/presentations/SMM/SMM.pdf. Accessed: 2021{18-06.
- Morten Nielsen. Performance measures. http://www.cbs.dtu.dk/courses/27625.algo/recordings/Performance_measure.mp4. Accessed: 2021{18-06.
- Morten Nielsen. Artificial neural network 1. http://www.cbs.dtu.dk/courses/27625.algo/presentations/NN-1/NNtalk_w_answers.pdf. Accessed: 2021{21-06.