Midterm meeting: Bachelor thesis of Robin Ender

Date: *2025-06-12*

Attendees: *Robin Ender, Asis Hallab, Eric Schumbera*

Current status of phase separation prediction

- Many Phase Separation Predictors have emerged, yet the performance is still not optimal
  - the best ones use scalar features (e.g. percentage of IDR)
- Problems are:
  - Lacking Training Data
  - Bias towards proteins containing IDRs
- Idea:
  - Using the sequence (block decomposition) to predict PS and try a newly published data set

Repetition - Block Decomposition

- The block decomposition algorithm takes a protein sequence and a mapping
- It outputs a list of blocks that all have a certain word balance (“uniformity”)
- Steps to use it for neural networks:
  - Adjusting the output
  - Finding relevant mappings

Adjusting the output of the block decomposition algorithm

The block decomposition algorithm was modified to yield an output that can be used in a neural network:

Old Output:
[(5, 14),(15, 22)...]

New Output:
[0,0,0,0,1,1,1,...,2,2,2...]

- The list of tuples was converted to one list representing the sequence.
- Labels were created, representing the most common group(s).

Finding meaningful mappings

Seven Mappings were found (in literature) that relate to phase separation or are generally meaningful:

Aliphatic - Aromatic - Positive - Negative
RG-Mappings (two separate)
IDR-Mapping
Most meaningful 5 mapping
PiPi-Mapping (two separate)

Those were used to generate one block decomposition for each mapping per protein.

Data sets I

For now two different data sets were used:

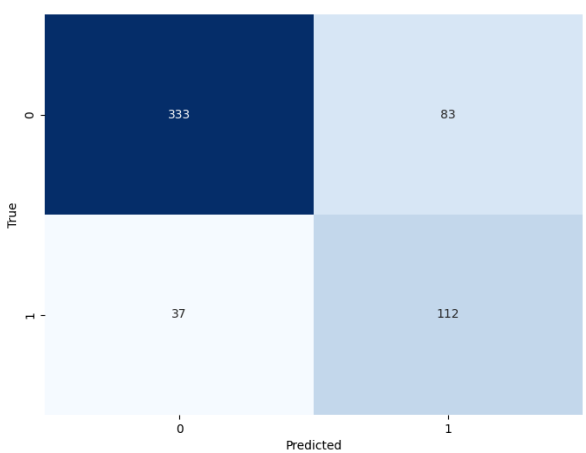
- PPMCLABs llps data set
  - created in an effort to create a dataset with an appropriate negative data set (many studies use RCSB Protein Data Bank, which are not guaranteed to be non phase separating)
  - contains 746 positive and 2077 negative entries

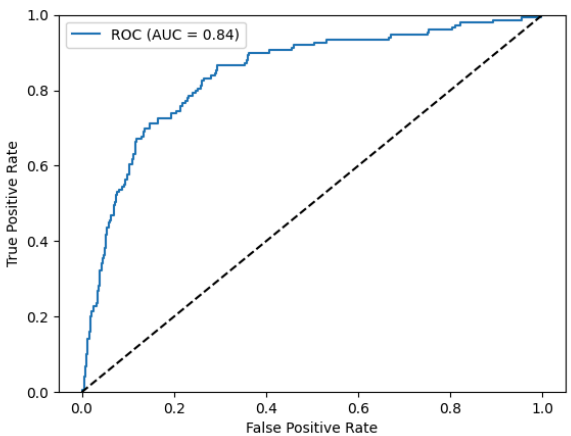
Data sets II

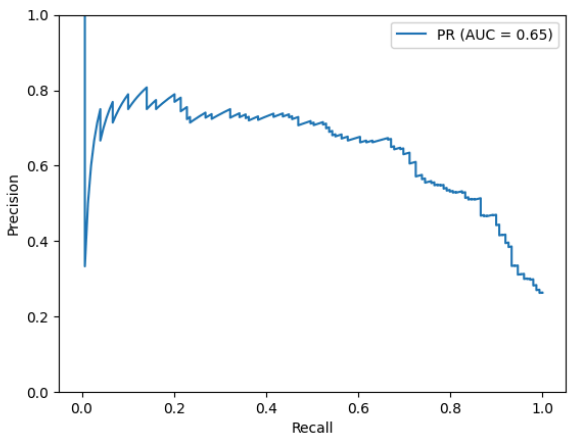
- PSPires data set
  - used to train PSPire, a recent well performing PS predictor addressing the difficulty of predicting phase separating proteins with no IDRs
  - contains 517 positive and around 10,000 negative entries

Test if a simple cnn model can learn from the data

To see if a cnn model was able to learn a basic 1 layer model was created and run on the PPMCLAB data set:

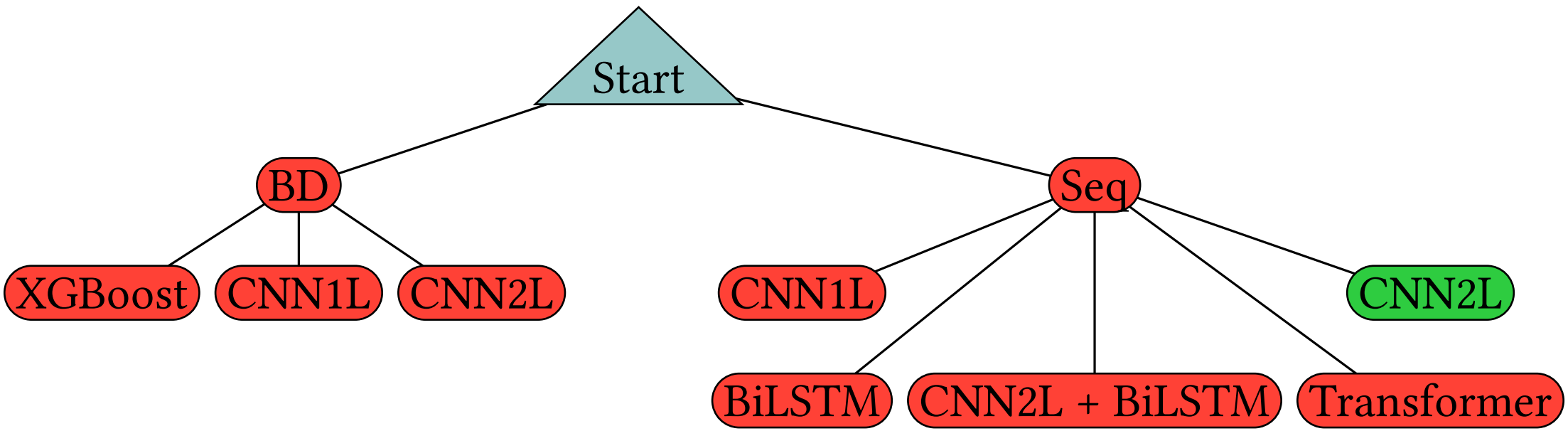




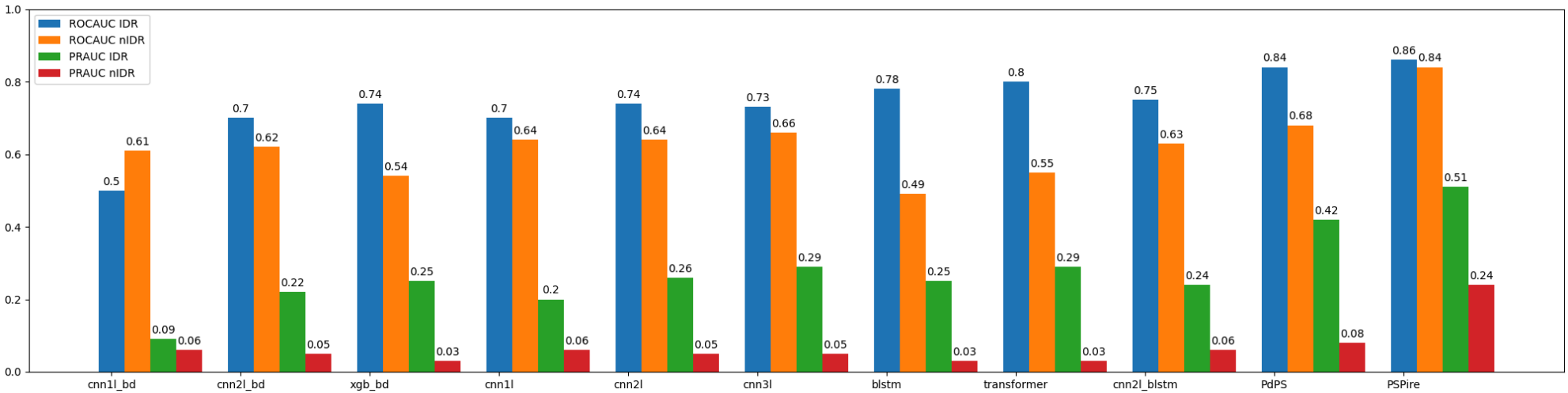


Testing different models

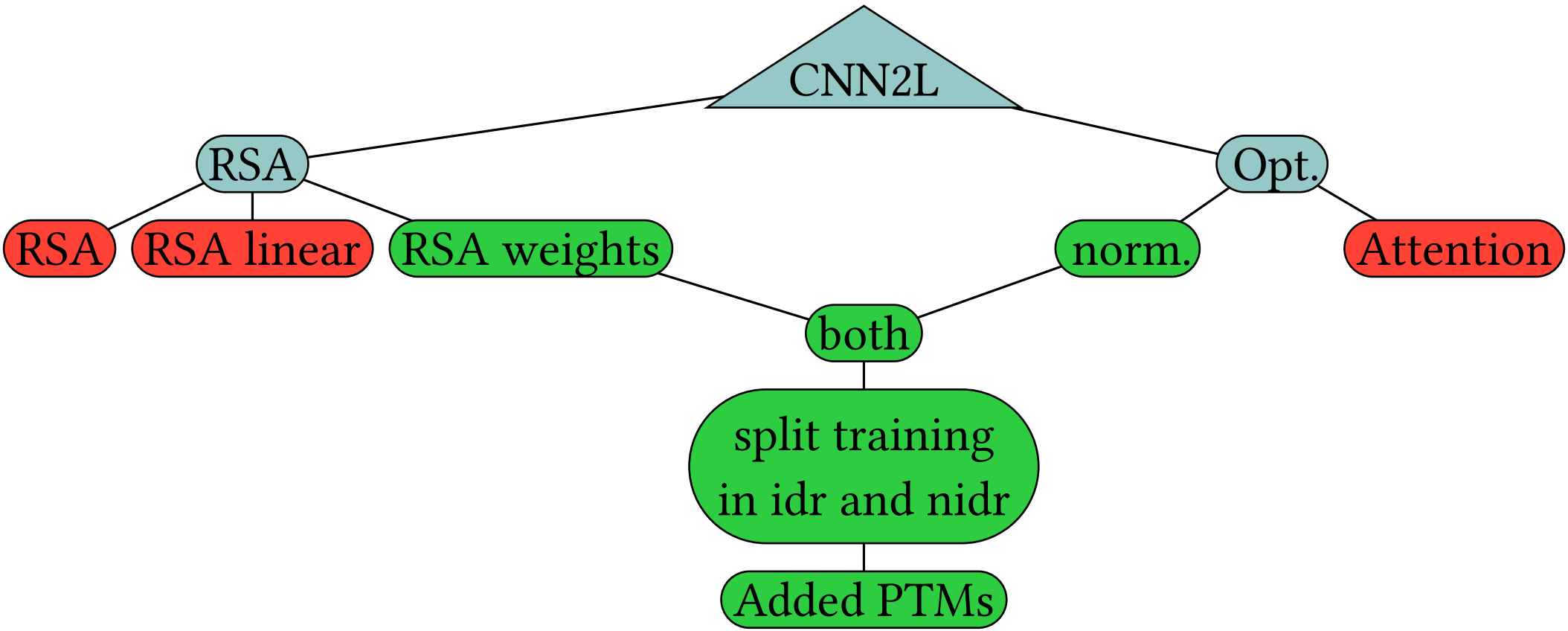
As the CNN2l model performed relatively good and was relatively quick in training it was used for further tests:



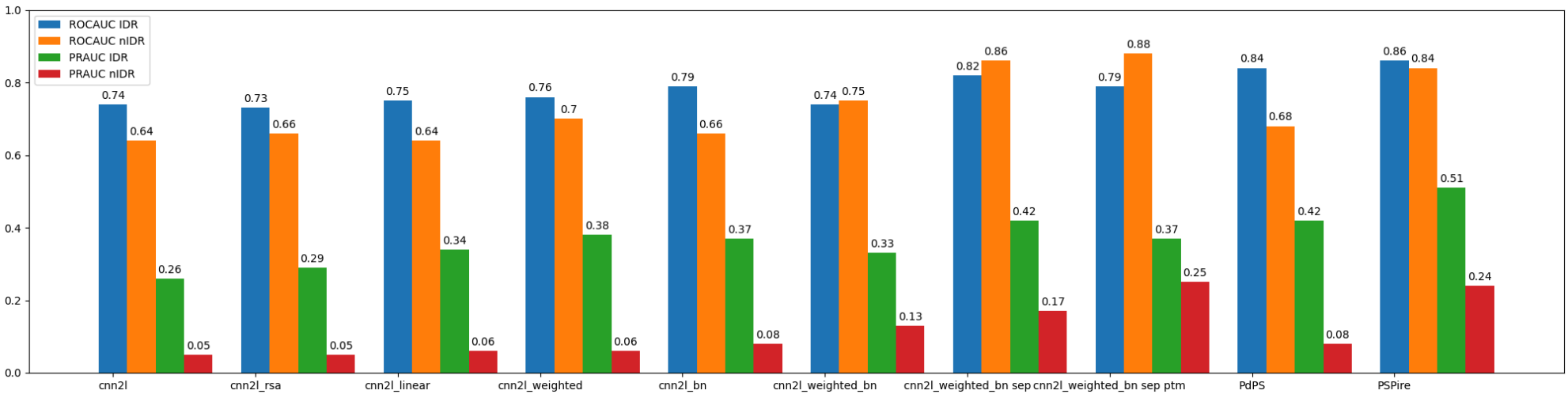
Performance of the initial models on PSPire data



Enhancing the CNN2L models



Performance of the new models



Performance on MLO data sets

PSPs	Dataset	Parameter	My model	PdPS	PSPire
noID-PSPs	G3BP1	ROCAUC	**0.96**	0.81	0.93
		PRAUC	0.51	0.18	**0.66**
	DACT1	ROCAUC	0.90	0.81	**0.93**
		PRAUC	0.49	0.18	**0.60**
	RNAGranule	ROCAUC	0.88	0.68	**0.90**
		PRAUC	0.18	0.08	**0.28**
	PhaSep	ROCAUC	**0.85**	0.65	0.80
		PRAUC	**0.73**	0.47	0.71
	DRLLPS	ROCAUC	0.80	0.68	**0.85**
		PRAUC	**0.77**	0.45	0.74

Performance on MLO data sets

PSPs	Dataset		My model	PdPS	PSPire
ID-PSPs	G3BP1	ROCAUC	0.74	0.86	**0.91**
		PRAUC	0.29	0.41	**0.58**
	DACT1	ROCAUC	0.72	0.85	**0.88**
		PRAUC	0.22	0.33	**0.35**
	RNAGranule	ROCAUC	0.80	0.82	**0.84**
		PRAUC	0.39	0.42	**0.48**
	PhaSep	ROCAUC	0.71	**0.74**	0.72
		PRAUC	0.70	**0.80**	0.79
	DRLLPS	ROCAUC	0.69	**0.76**	0.75
		PRAUC	0.71	0.77	**0.78**

Results in context of the initial idea of this work

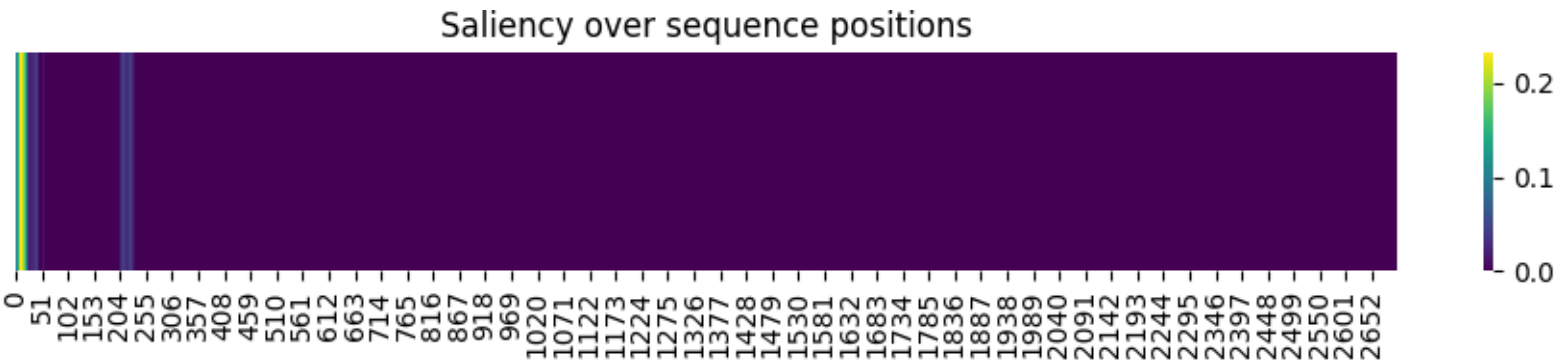
- block decomposition can be used to predict phase separation
  - but using the raw sequence yields better results
  - therefore the focus has shifted to using the sequence and additional data to build a good phase separation predictor
- while the ppmclab data set should be a beeter training set, it failed to predict the MLOs

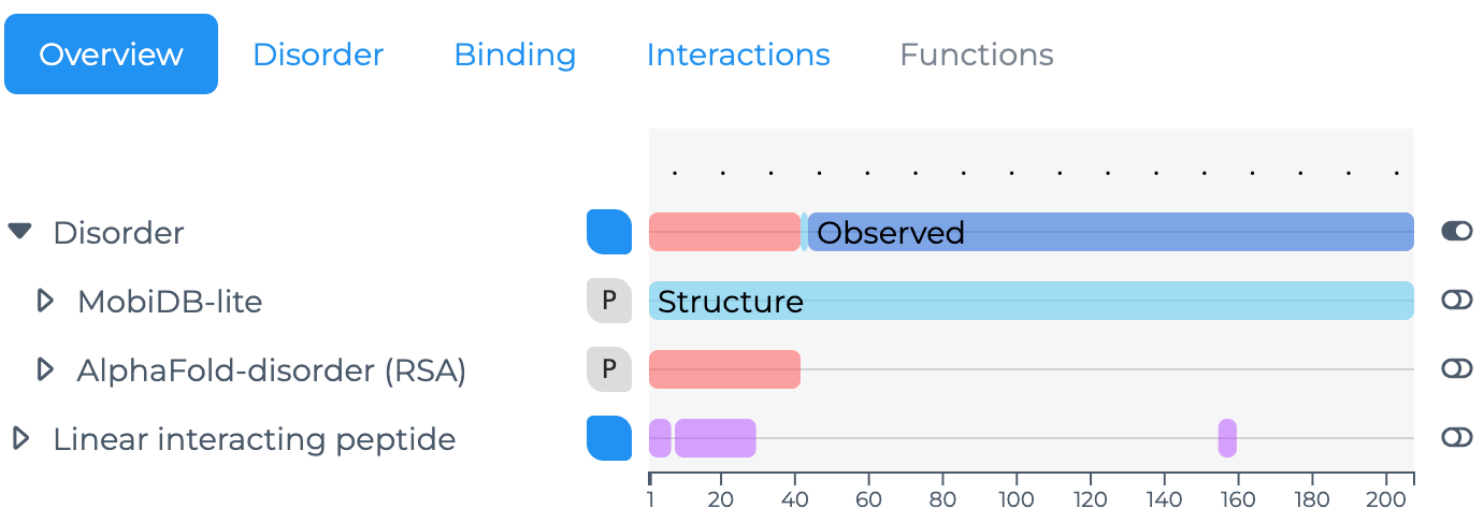
Visualizing the results

- ROC AUC and PR AUC
- bar plots / tables for comparison with other tools
- Saliency (shows which input positions the model “cares about most” when making its prediction) probably only for some visualizations

https://bbb.rlp.net/rooms/hal-cta-mfd-wzm/join

Saliency


Saliency over sequence positions
0.2
0.1
0.0
0
51
102
153
204
255
306
357
408
459
510
561
612
663
714
765
816
867
918
969
1020
1071
1122
1173
1224
1275
1326
1377
1428
1479
1530
1581
1632
1683
1734
1785
1836
1887
1938
1989
2040
2091
2142
2193
2244
2295
2346
2397
2448
2499
2550
2601
2652


Overview
Disorder
Binding
Interactions
Functions
Disorder
MobiDB-lite
AlphaFold-disorder (RSA)
Linear interacting peptide
Observed
Structure
P
P
1
20
40
60
80
100
120
140
160
180
200

What can / should be done in the remaining time?

- Cross validation of the final model
- writing!
- next week, report will be send
- last meeting: 10.07.
- end of this thesis: 04.08.
- personal deadline: 25.07.

Structure of the thesis

- Introduction
  - Liquid-Liquid Phase Separation
  - Block decomposition of protein sequences
  - Current predictors and the difficulties
  - Machine Learning in Bioinformatics
    - CNN
- Material
  - Data (explain and visualize data sets)
- Methods
  - Tools

- data preparation
- model architectures
- model optimizations

- Results
  - comparison of my own models
    - most important block decompositions
  - comparison with other models
    - on their test data
    - on mlo data
  - visualizations of important sequence segments

- Discussion

- Usefulness of this model
- What should / could still be done