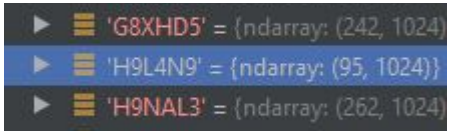
Binding Residue Prediction

Final Presentation of Group 1

Georg Böhm, Christian Hoffmann, Isabell Orlishausen, Leon Schwartz

June 24, 2021

Recap

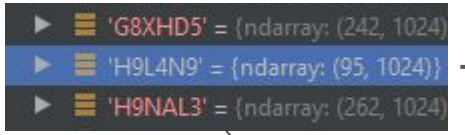


binary classification

```
>H9L4N9
MQINIQGHHIDLT...
0000100110001...
```

for each residue:
binding/non-binding

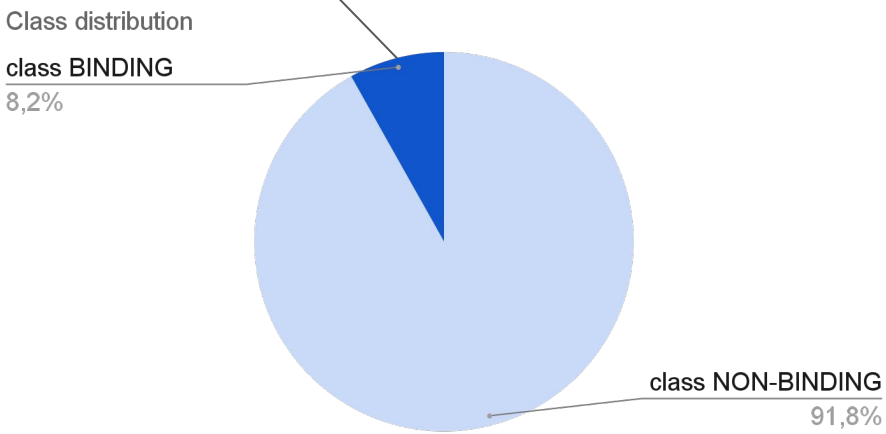
Recap



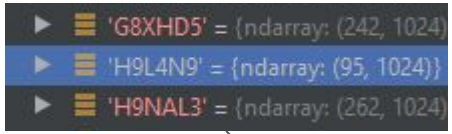
binary classification

```
>H9L4N9
MQINIQGHHIDLT...
0000100110001...
```

for each residue:
binding/non-binding



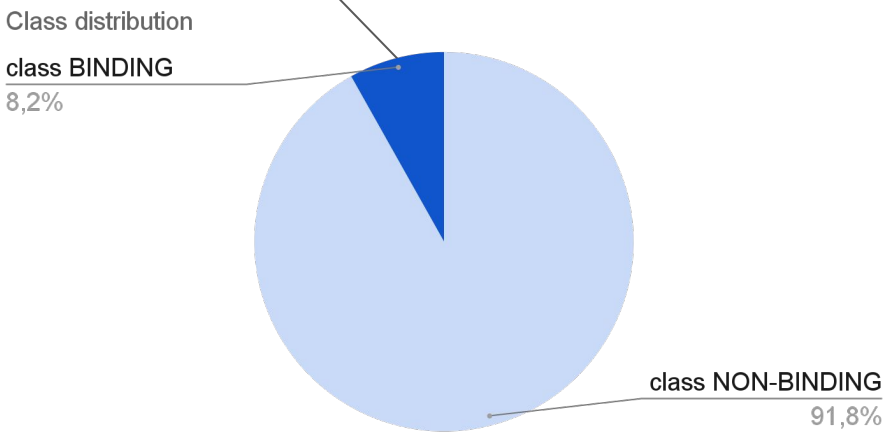
Recap



binary classification

```
>H9L4N9
MQINIQGHHIDLT...
0000100110001...
```

for each residue:
binding/non-binding



- manually created splits for CV
- common test set:
  - 20% of data, meaning
  - **34350** residues from
  - 195 proteins

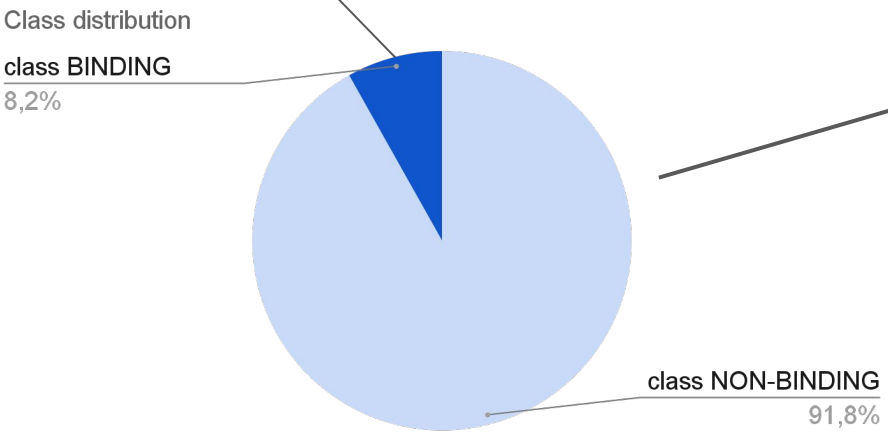
Recap

'G8XHD5' = {ndarray: (242, 1024)
'H9L4N9' = {ndarray: (95, 1024)}
'H9NAL3' = {ndarray: (262, 1024)

binary classification

```
>H9L4N9
MQINIQGHHIDLT...
0000100110001...
```

for each residue:
binding/non-binding



Model scoring based on MCC:

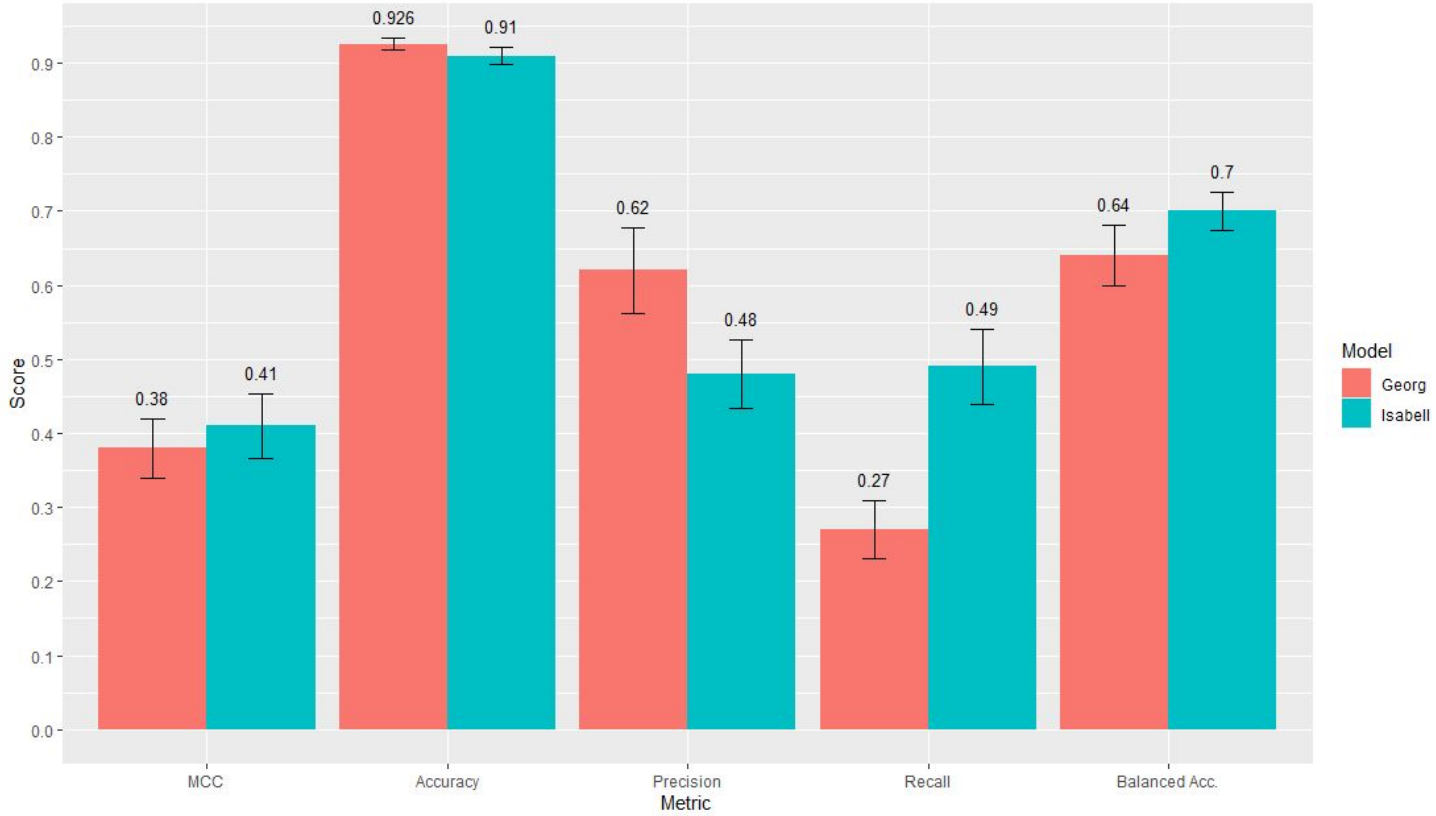
$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Recap - MSA embedding types

MSA embedding	MSA1	MSA2	MSA3
Characterization	generated with default parameters	restricted to generally contain less sequences	contains only sequences that have a sequence similarity of at least 0.5

Comparison of best models - performance

comparing **mean** validation scores assessed on **training** set



Comparison of best models - parameters

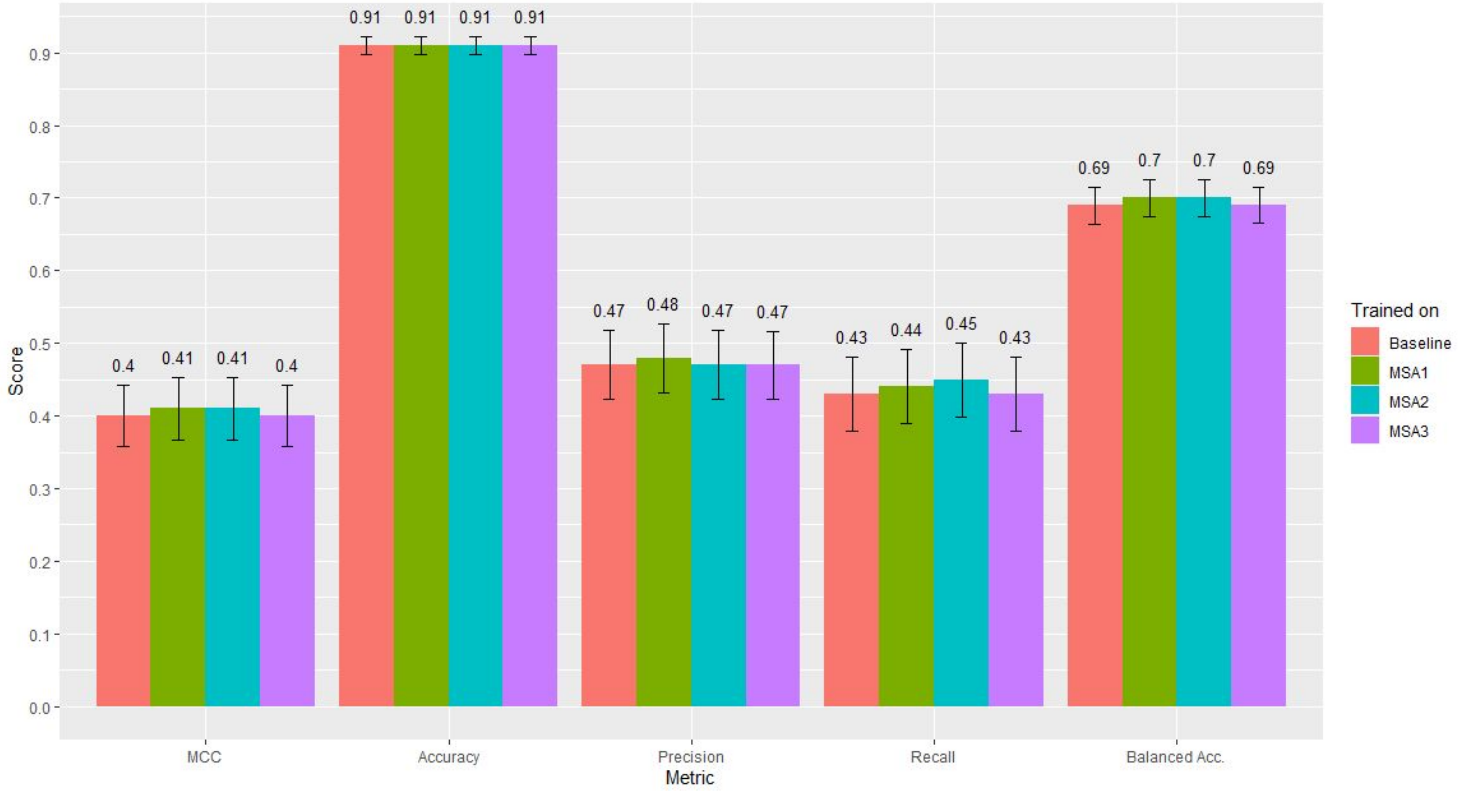
Parameter/Category	Model Isabell: MLP Classifier	Model Georg: MLP Classifier
activation, solver, learning_rate_init	same defaults (‘relu’, ‘adam’, 0.001)	same defaults (‘relu’, ‘adam’, 0.001)
early_stopping	True	True
alpha	0.001	0.0001
hidden_layer_sizes	(80,)	(100, 50)
learning_rate	'invscaling'	'adaptive'
max_iter	50	175
Resampling	Random Undersampling (0.2)	-

→ Found best set of **parameters**!

But: Which **embedding type** should be used for training?

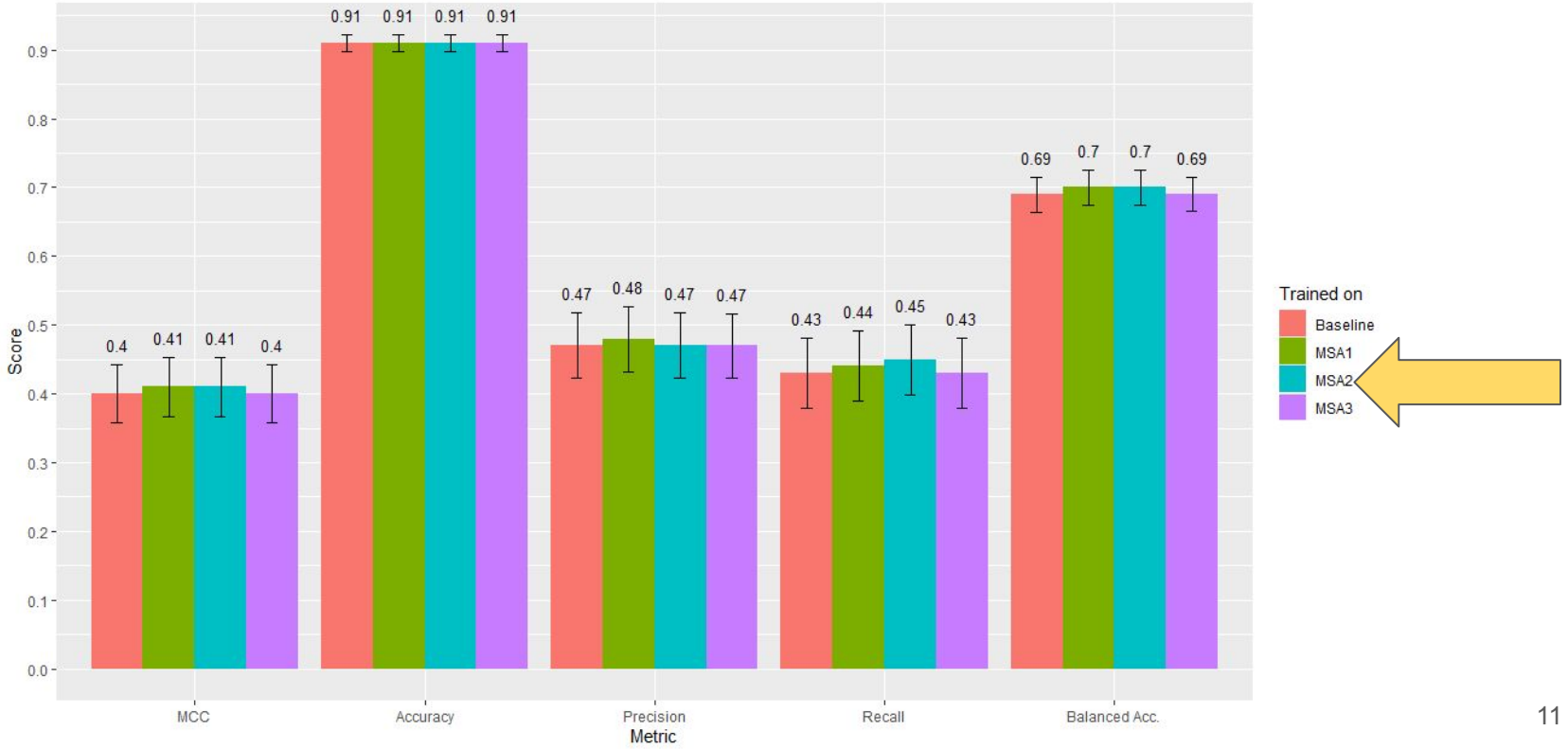
Comparison of training performance with final params

comparing **mean** validation scores assessed on **training** set



Comparison of training performance with final params

comparing **mean** validation scores assessed on **training** set

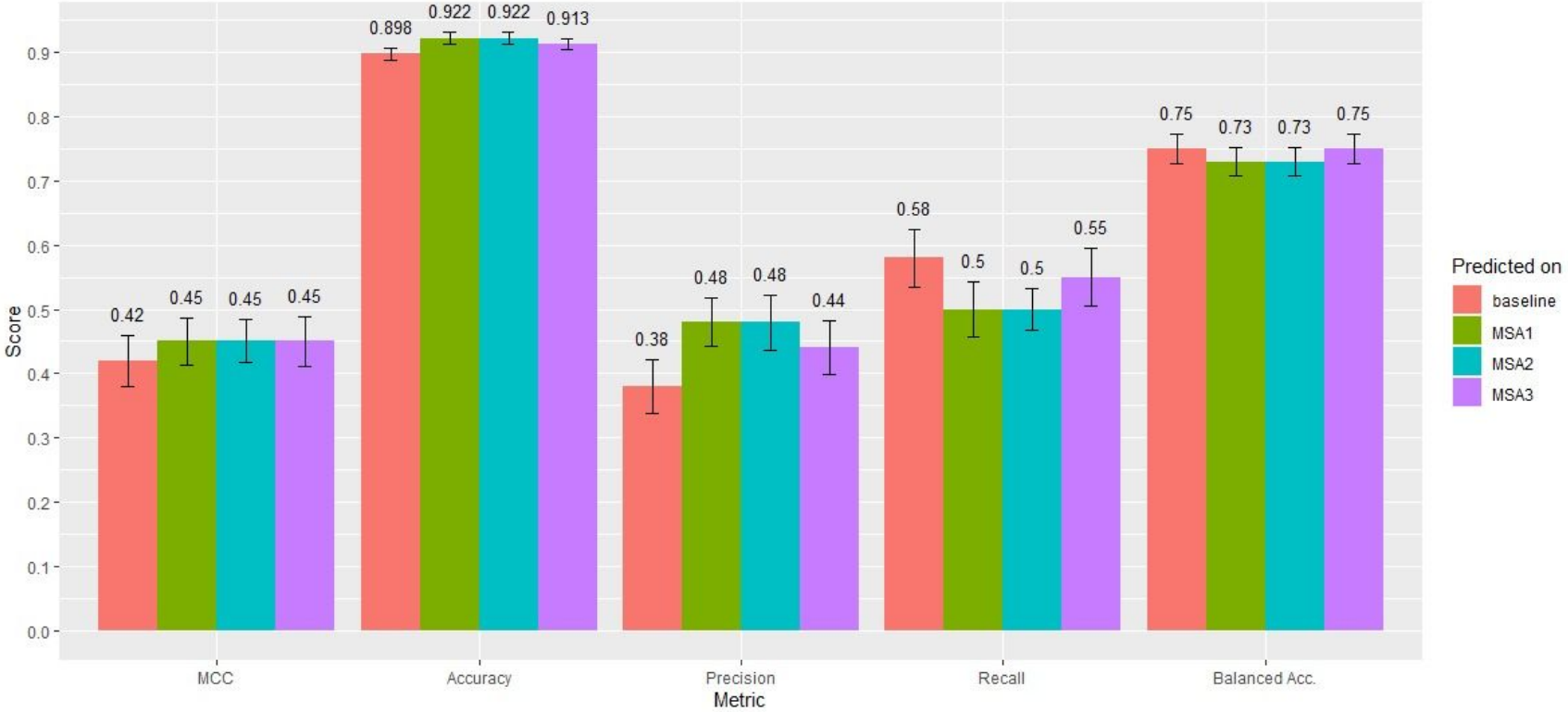


Final model

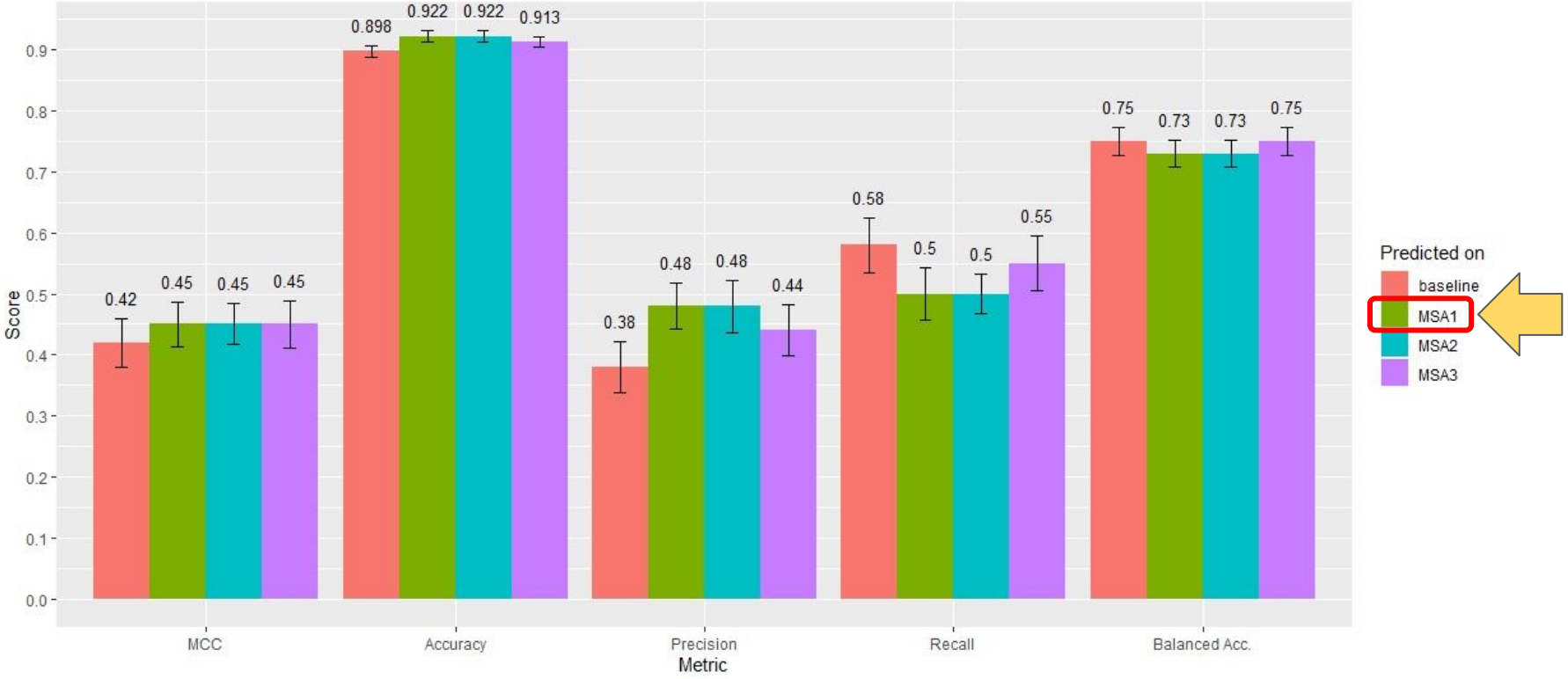
Parameter/Category	Model Isabell: MLP Classifier	Model Georg: MLP Classifier
activation, solver, learning_rate_init	same defaults (‘relu’, ‘adam’, 0.001)	same defaults (‘relu’, ‘adam’, 0.001)
early_stopping	True	True
alpha	0.001	0.0001
hidden_layer_sizes	(80,)	(100, 50)
learning_rate	'invscaling'	'adaptive'
max_iter	50	175
Resampling	Random Undersampling (0.2)	-

… and using **MSA2 embeddings** for training!

Final model - performance

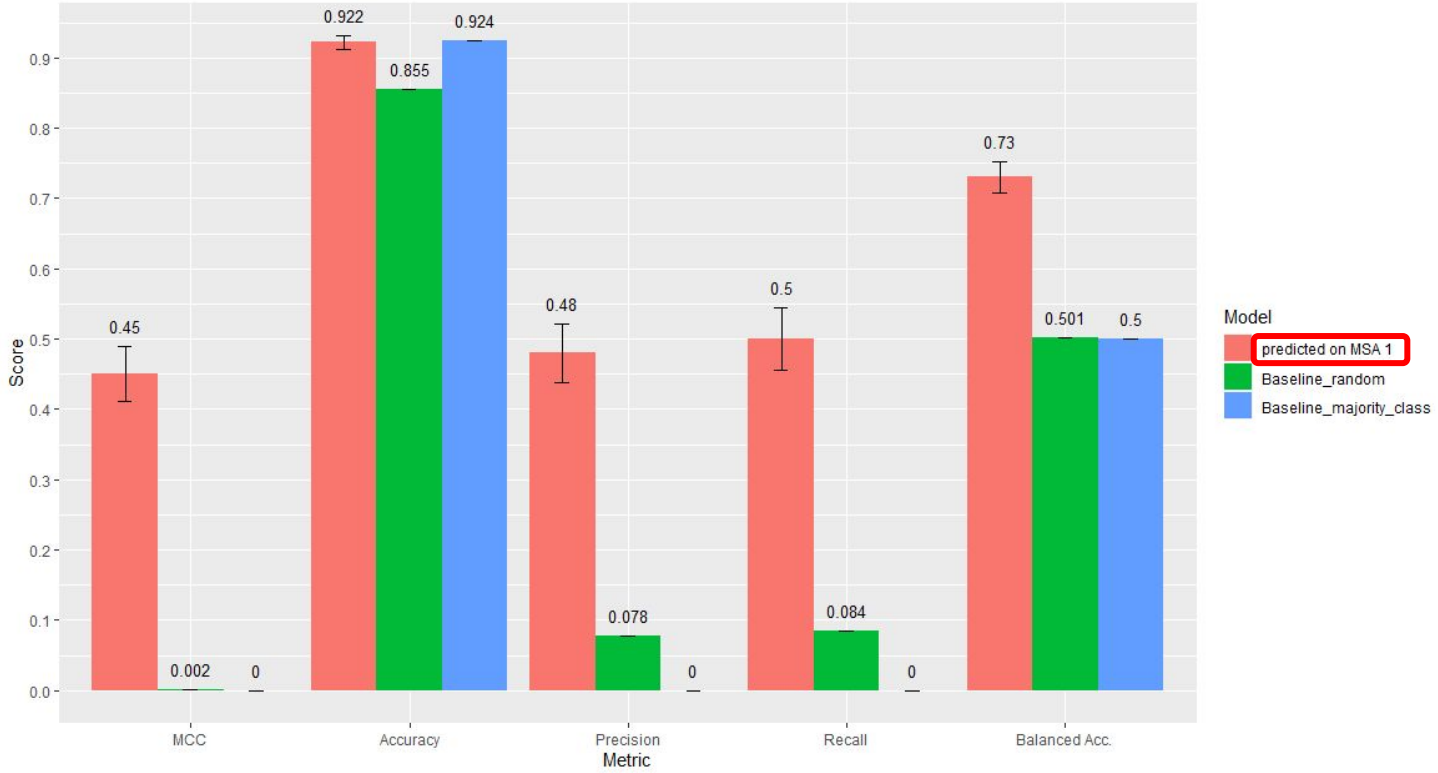


Final model - performance



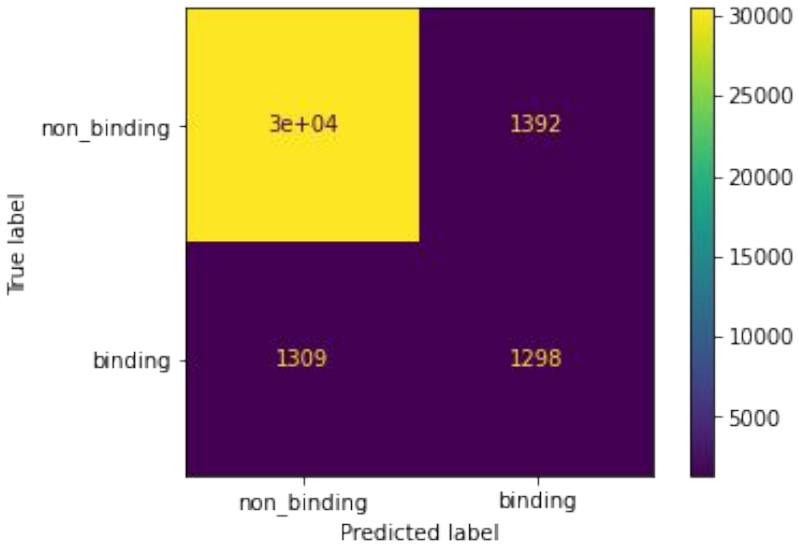
Final model - performance

comparing **prediction** scores assessed on **test** set

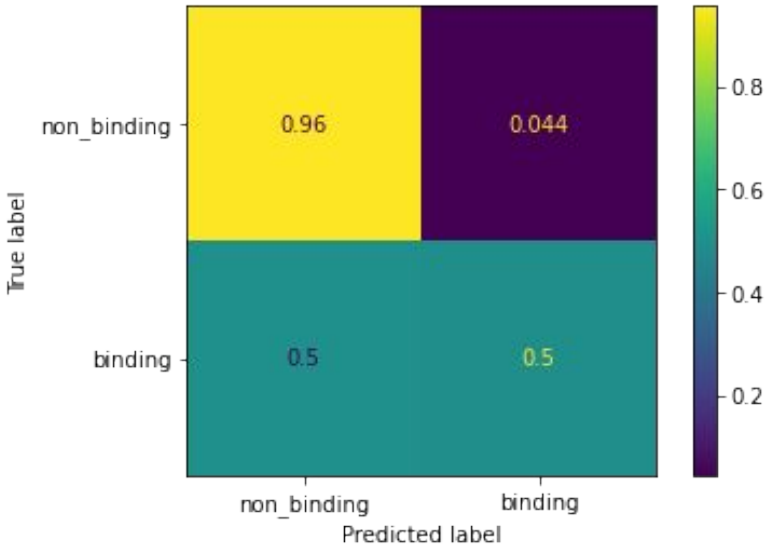


Final model - performance

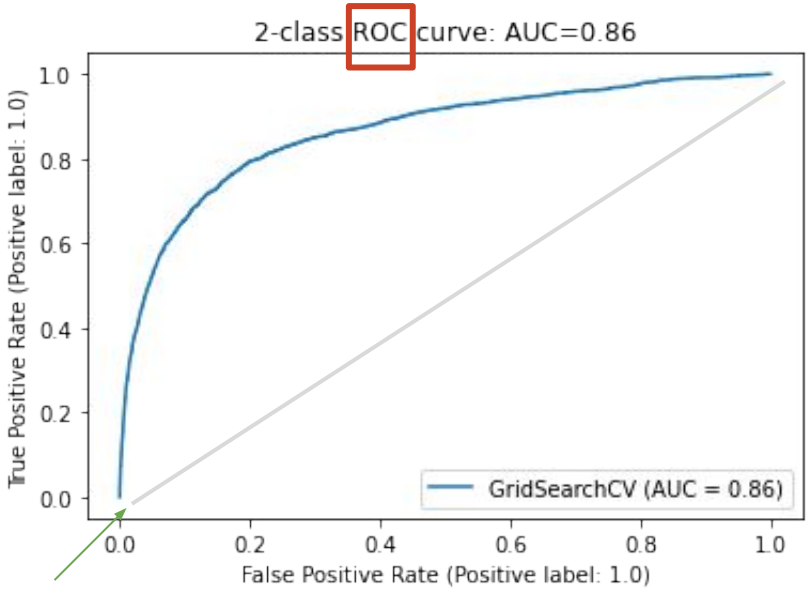
non-normalized



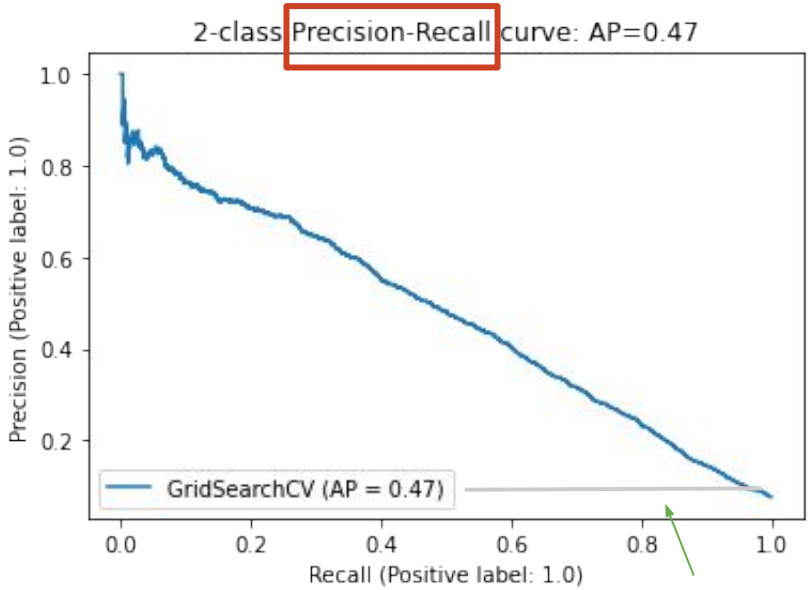
normalized



Final model - performance



random classifier



random classifier

Some “interesting” findings during model development

On MSA1 embeddings (Isabell):

- oversampling: **improved** MCC for **RFC**, but worsened it for MLP
- undersampling: **improved** MCC for **MLP**, but worsened it for RFC
- tried more sophisticated **undersampling** techniques, no improvement
- always precision >> recall, until increasing RUS rate

On MSA3 embeddings (Georg):

- multiple hidden layers: **improved** prediction scores of MSA-3 Model
- undersampling not carried out for this model, still reasonable scores compared to final model

Thank you for listening!