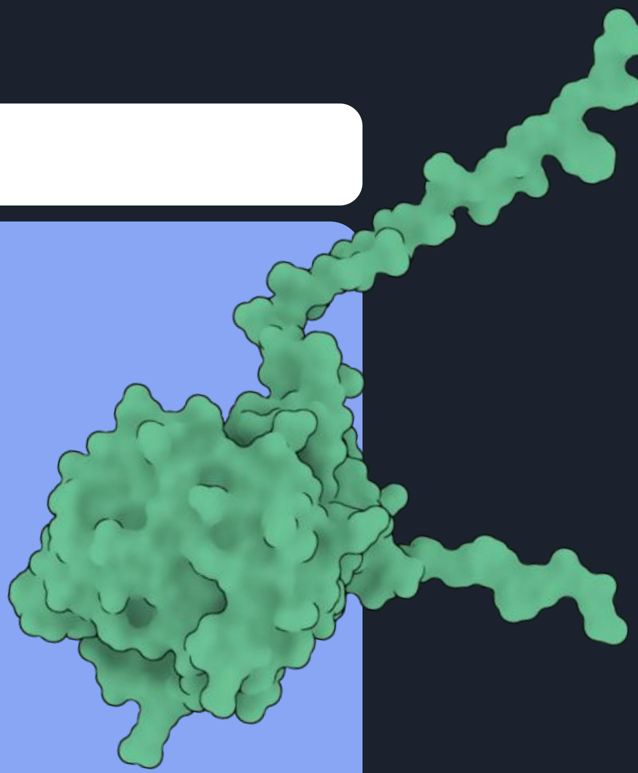
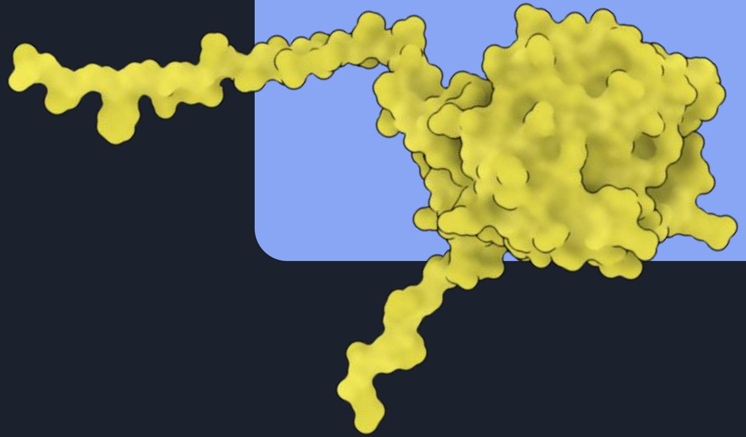


Rab Protein Classification

Gaëtan, Luca, Sydney

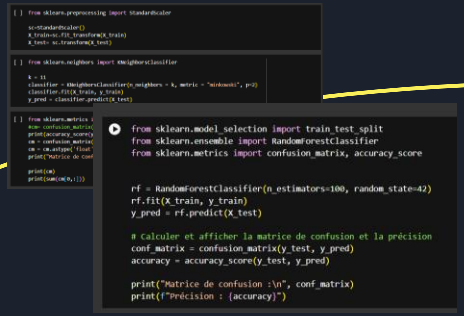
Mentor: Deepak Yadav



Email: gaetan.ray@unil.ch, luca.paley@unil.ch, sydney.fleming@unil.ch

Goal of our project

Classify the Rabs into their group

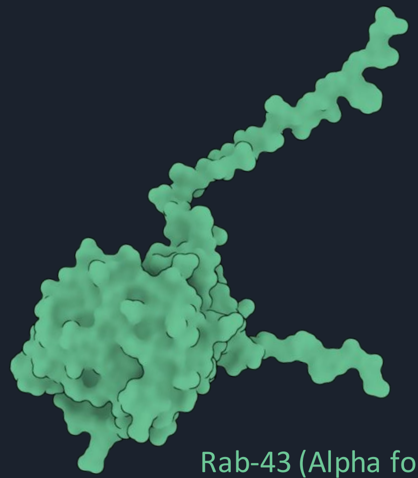


```
[ ] from sklearn.preprocessing import StandardScaler

    sc=StandardScaler()
    X_train=sc.fit_transform(X_train)
    X_test= sc.transform(X_test)
```

```
[ ] from sklearn.neighbors import KNeighborsClassifier

    k = 11
    classifier = KNeighborsClassifier(n_neighbors = k, metric = "minkowski", p=2)
    classifier.fit(X_train, y_train)
    y_pred = classifier.predict(X_test)
```

```
[ ] from sklearn.metrics i
    #cm= confusion_matrix(
    print(accuracy_score(y
    cm = confusion_matrix(
    cm = cm.astype('float'
    print("Matrice de conf

    print(cm)
    print(sum(cm[0,:]))
```

```
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import confusion_matrix, accuracy_score


rf = RandomForestClassifier(n_estimators=100, random_state=42)
rf.fit(X_train, y_train)
y_pred = rf.predict(X_test)

# Calculer et afficher la matrice de confusion et la précision
conf_matrix = confusion_matrix(y_test, y_pred)
accuracy = accuracy_score(y_test, y_pred)

print("Matrice de confusion :\n", conf_matrix)
print(f"Précision : {accuracy}")
```



Rab-43 (Alpha fold)

Group 1

Group 2

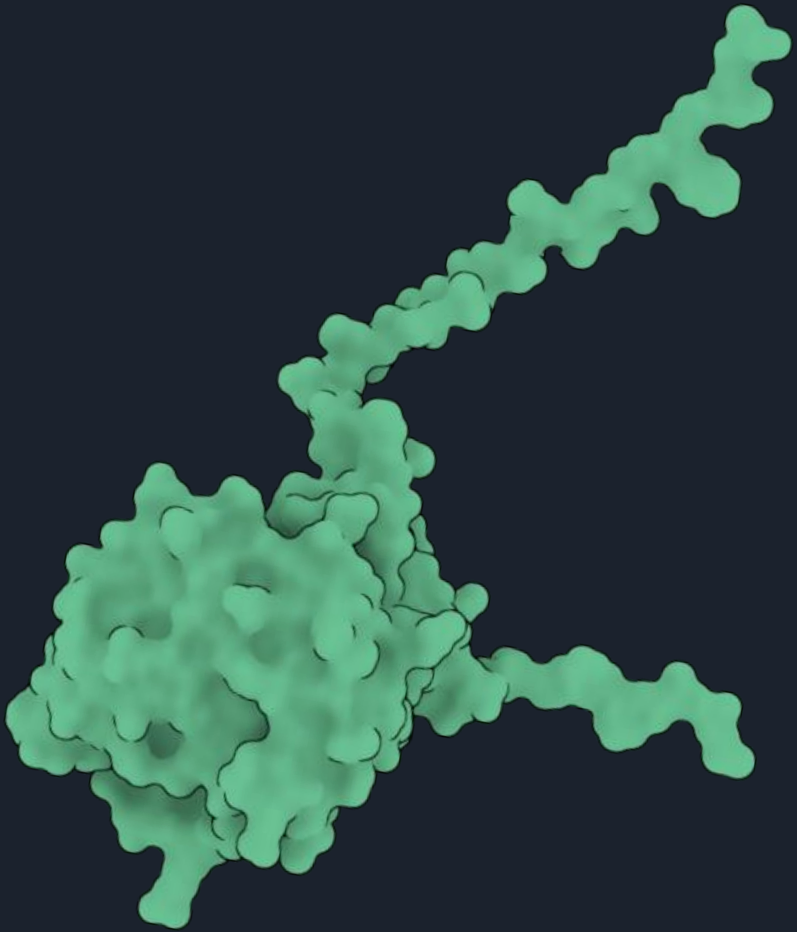
Group 3

Group 4

Group 5

Group 6

What are Rab proteins?



Rab-43 (Alpha fold)

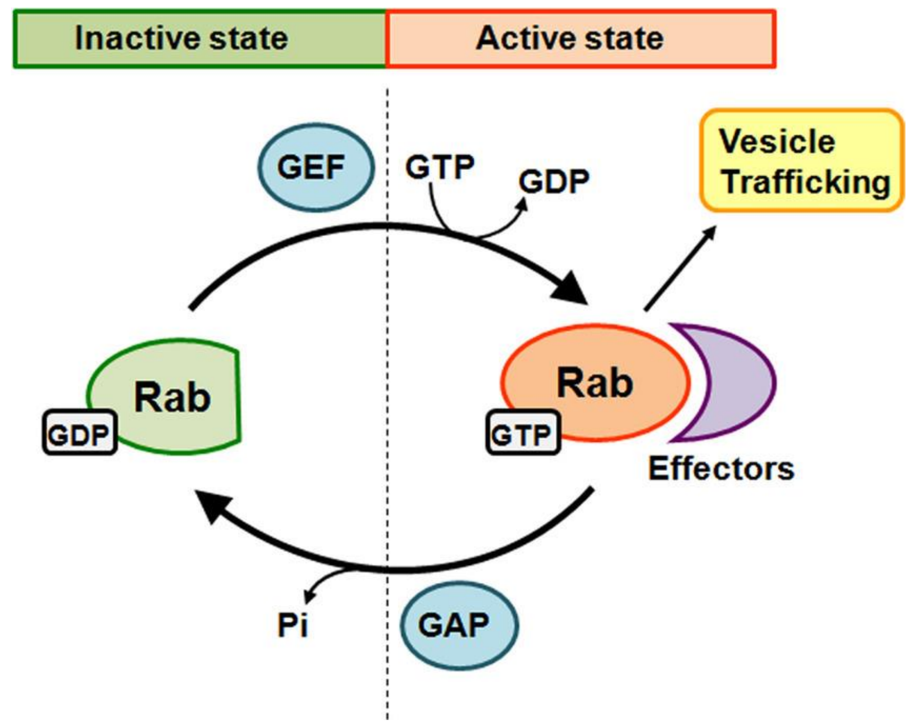
What are Rab proteins?

GTPase with functions in tethering in **vesicular trafficking**

David Goodsell

What are Rab proteins?

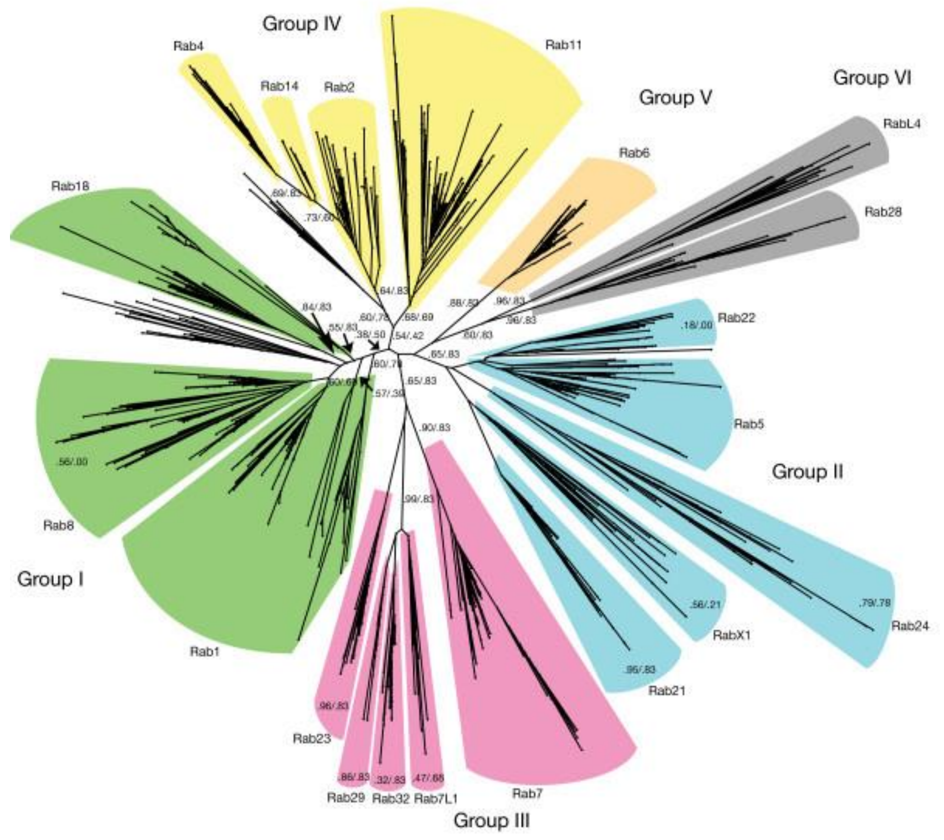
GTPase →
exchange GDP against
GTP to activate effectors



Wang S. *et al*, 2017

What are Rab proteins?

Large **family** of proteins



Klöpper, T.H. *et al.* 2012

… and why study them?

Griscelli syndrome

- partial albinism
- immunological defects
- primary neurological dysfunctions

Choroideremia

- Blindness

Neurological disorders

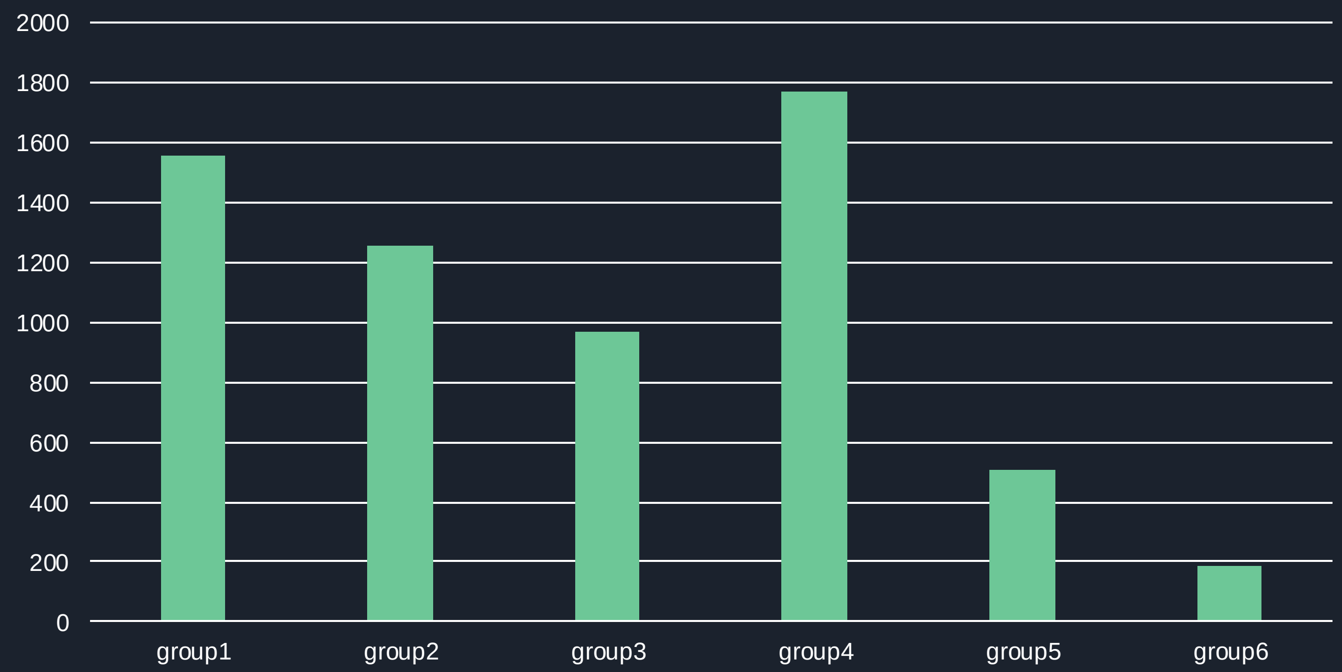
Rab-associated **diseases**

Pipeline and methods

What does our data look like?

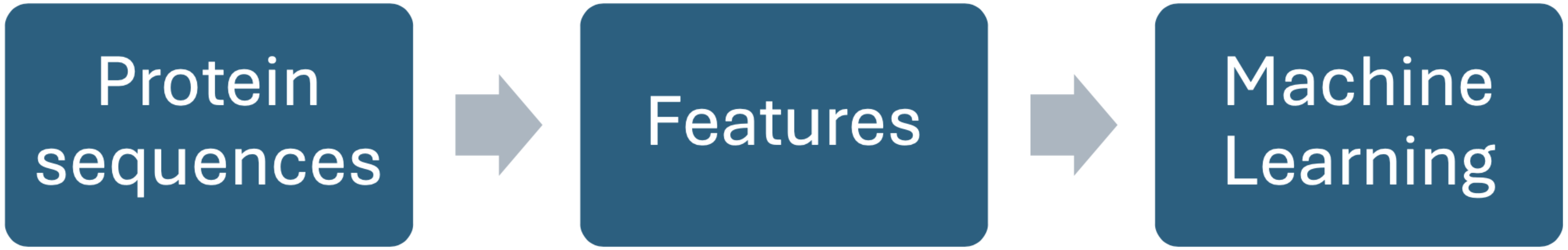
```
>PoTr_Rab1C
MNPEYDYLFKLLLIGDSGVGKSCLLLRFADDSYLESYISTI
GVDFKIRTVEQDGKTIKLQIWDTAGQERFRTITSSYYRGAH
GIIVVYDVTDQESFNNVKQWLNEIDRYASENVNKLLVGNKC
DLTANKVVSYETAKAFADEIGIPFMETSAKNATNVEQAFMA
MAAEIKNRMASQPAMNNARPPTVQIRGQPVNQKSGCCSS
>PaTe_Rab1_1
MSLQQEYDYLFKILLIGNSAVGKSSLLLRFADNVFNESFLP
TIGVDFKIRTFDLNGKTVKLQIWDTAGQERFKTITNSYYKG
AHGIILVYDVTDKQSFKDVENWLAEVEKYANENVVRVLVGN
KVDLESKREVTSEEGKELADSLNIRFIETSAKNSSNVEKAF
ITLANEIKAKVAKSSEAIPVKTGPRITPDQQQNTVKDTGCC
>BrMa_Rab1
MVSINPEYDYLFKLLLIGDSGVGKSCLLLRFADDTYTESYI
STIGVDFKIRTIDLNGKTIKLQIWDTAGQERFRTITSSYYR
GAHGIIVVYDITDQESFNNVKQWLQEIDRYACENVNKLLVG
NKCDLIIRRAVEHSAAKEYADQLGIPFLETSAKSSTNVEQA
FLTMASEIKNRMGPIQQVGTGPSVRIGGSQPVNEKKSGGCC
>CaAl_Ypt1
MNNEYDYLFKLLLIGDSGVGKSCLLLRFADDTYTPDYISTI
GVDFKIRTIELDGKTIKLQIWDTAGQERFRTITSSYYRGAH
GIIIVYDVTDQESFNNVKQWLQEIDRYATGGVMKLLVGNKA
DLSDKKIVEYTAAKEFADALDIPFLETSALSSTNVEQAFYT
MARQIKAQMTNNANAGNAANAKGKSNVNLRGESLTSNQSNS
CC
>PaTe_Rab1
MIKEYDYLFKLVIIGNSGVGKSSLLLRFADDQFSESYLTTI
GVDFRFRTLPIDGKNVKLQIWDTAGQERFRTITSAYYKGAD
GIVMVYDVTQGQSFDDIDKFWLHEVESYGEKNVQLLIIGNK
NDLDEQKQVETSKAEEYCKSHNMLFMECSAKTADHVNNAFL
ELSRKLMAKKDASQPPKTTNTTSNASQQSQSRGQTNTNTQQ
SKQLSAGNTNQKKQKDGGCC
```

Number of Rab proteins per group



Unequal distribution

What did we do?



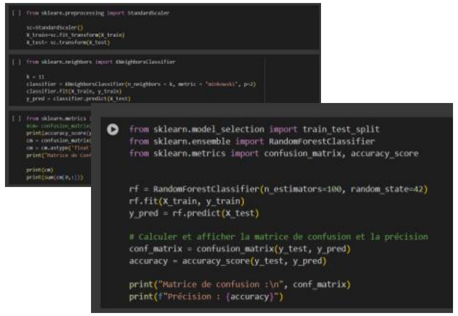
Protein sequences

Features

Machine Learning

>RAB 1
KVQKLSKTVIEVASLQNELATHL
HVQTQNINVLLD
>RAB 2
KAEALRRTAVDIASLQSELATHL
QVQTQNIN
>RAB 3
EVRQIEGRVVEISRLQEIFTEKVL
QQEAEIDSIHQ

Seq	Feat1	Feat2	Feat3
RAB 1	0.87	0.65	0.23
RAB 2	0.03	0.12	0.54
RAB 3	0.34	0.65	0.49



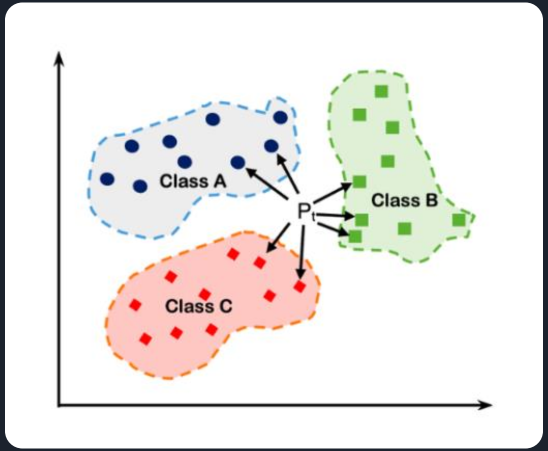
```
from sklearn.preprocessing import StandardScaler

sc=StandardScaler()
X_train=sc.fit_transform(X_train)
X_test= sc.transform(X_test)
```

```
from sklearn.neighbors import KNeighborsClassifier

k = 11
classifier = KNeighborsClassifier(n_neighbors = k, metric = "minkowski", p=2)
classifier.fit(X_train, y_train)
y_pred = classifier.predict(X_test)
```

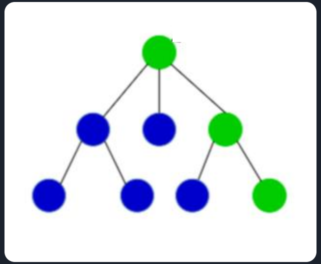
```
from sklearn.metrics 
#cm= confusion_matrix(
print(accuracy_score(y
cm = confusion_matrix(
cm = cm.astype('float'
print("Matrice de Conf

print(cm)
print(sum(cm[0,:]))
```

```
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import confusion_matrix, accuracy_score


rf = RandomForestClassifier(n_estimators=100, random_state=42)
rf.fit(X_train, y_train)
y_pred = rf.predict(X_test)

# Calculer et afficher la matrice de confusion et la précision
conf_matrix = confusion_matrix(y_test, y_pred)
accuracy = accuracy_score(y_test, y_pred)

print("Matrice de confusion :\n", conf_matrix)
print(f"Précision : {accuracy}")
```

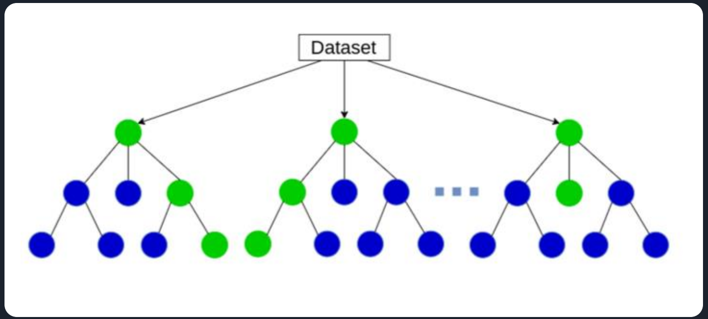
Methods used



KNN



Decision tree



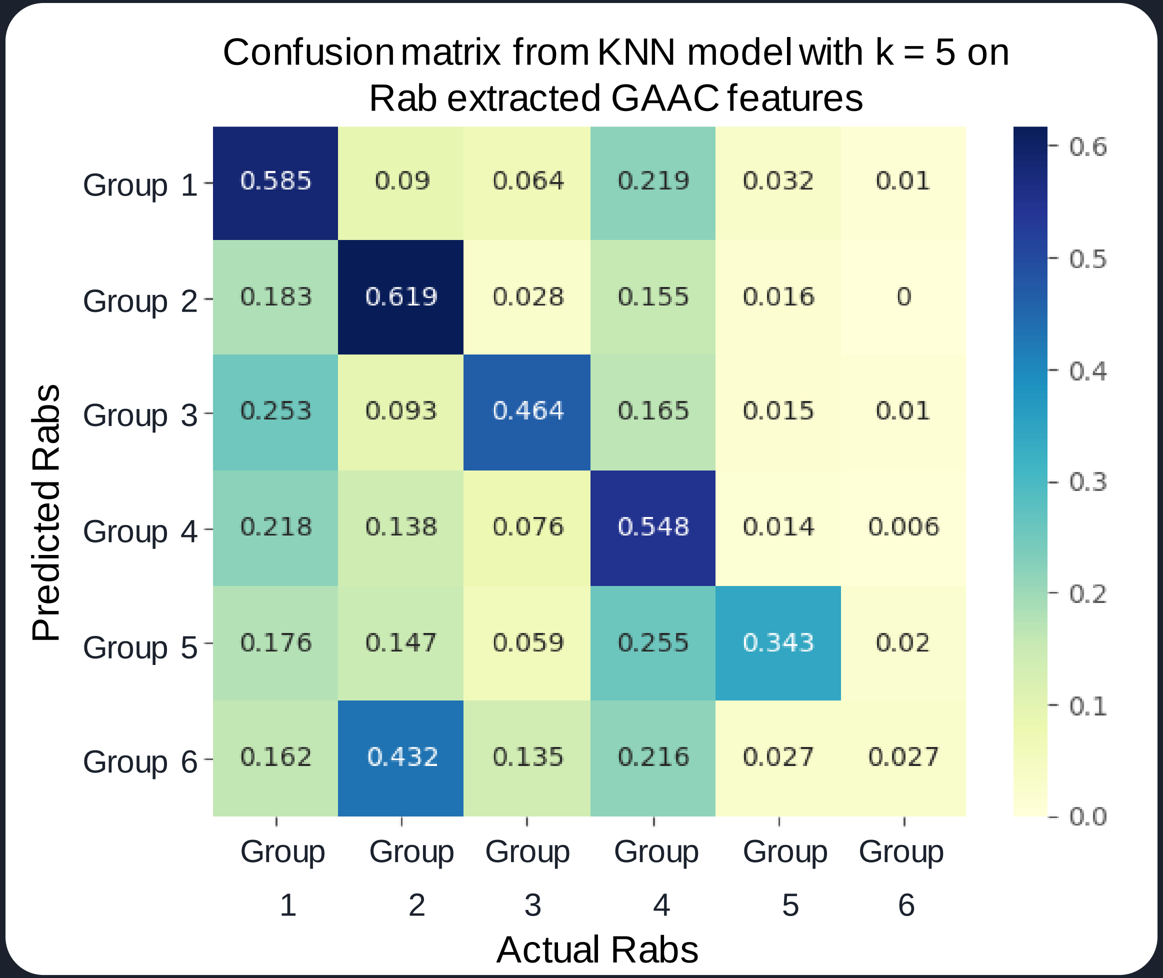
Random forest

First feature: GAAC

- GAAC = grouped amino acid composition
- **Physico-chemical** properties

	Aliphatic	Aromatic	Positively charged	Negatively charged	Uncharged
rab1	0,2	0,3	0,1	0,3	0,1
rab2	0,3	0,1	0,3	0,1	0,2
rab3	0,1	0,2	0,3	0,3	0,1

GAAC predictor using **KNN**



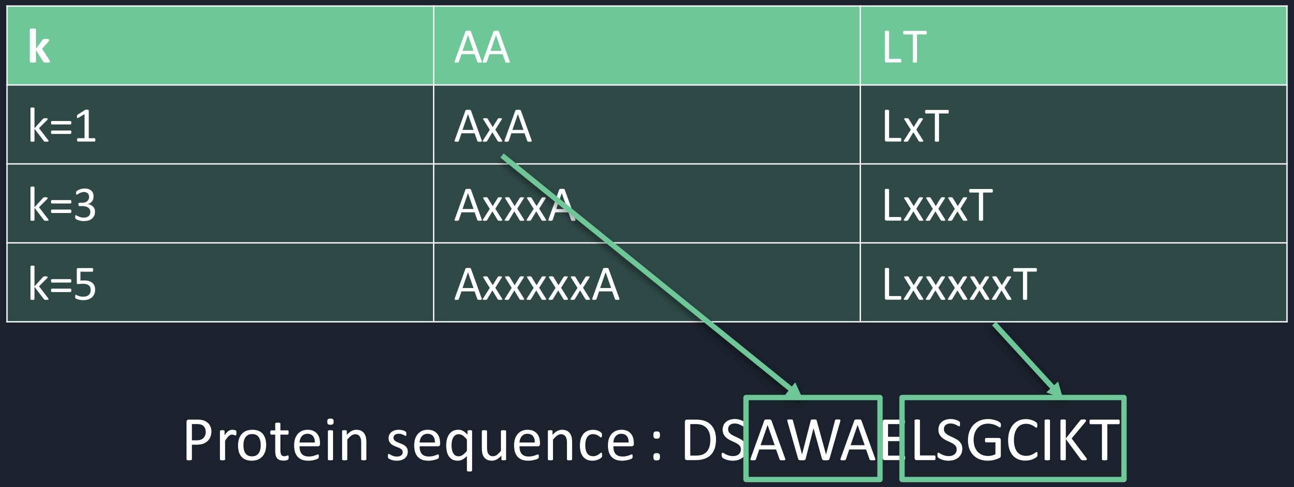
Confusion matrix from KNN model with k = 5 on Rab extracted GAAC features

	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
Group 1	0.585	0.09	0.064	0.219	0.032	0.01
Group 2	0.183	0.619	0.028	0.155	0.016	0
Group 3	0.253	0.093	0.464	0.165	0.015	0.01
Group 4	0.218	0.138	0.076	0.548	0.014	0.006
Group 5	0.176	0.147	0.059	0.255	0.343	0.02
Group 6	0.162	0.432	0.135	0.216	0.027	0.027

Accuracy: 53%
Precision: 0.523
Recall: 0.526
F1-Score: 0.519

Second feature: CKSAAP

- CKSAAP = Composition of **k-spaced** amino acid pairs

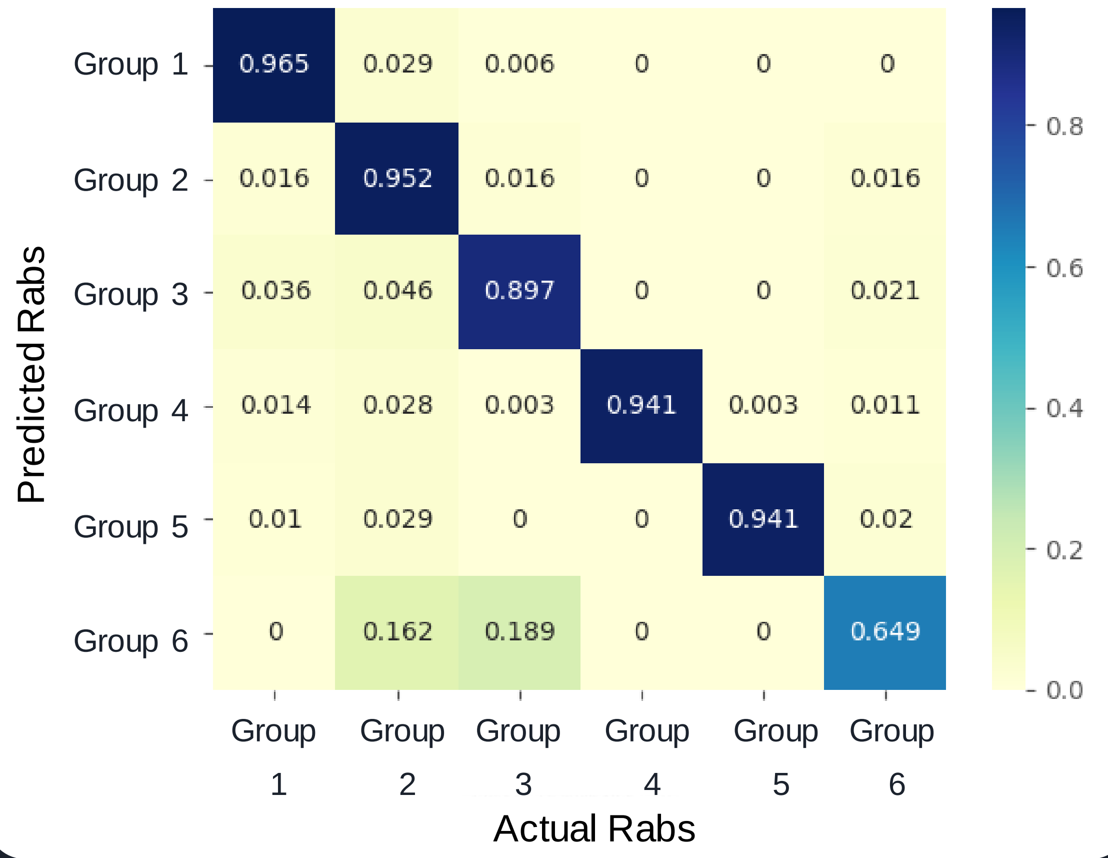


k	AA	LT
k=1	AxA	LxT
k=3	AxxxA	LxxxT
k=5	AxxxxxA	LxxxxxT

Protein sequence : DSAWAELSGCIKT

CKSAAP predictor using KNN

Confusion matrix from KNN model with k = 5 on Rab extracted CKSAAP features



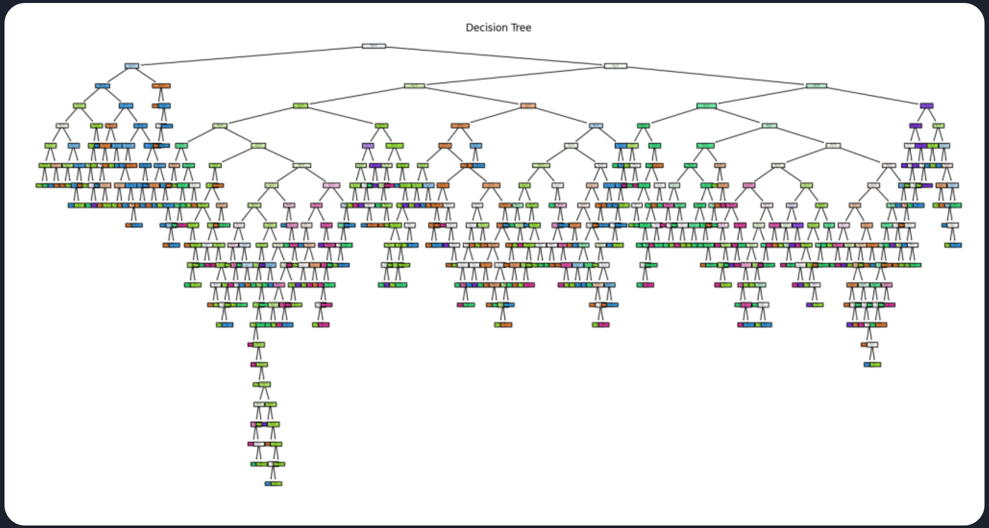
Predicted Rabs \ Actual Rabs	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
Group 1	0.965	0.029	0.006	0	0	0
Group 2	0.016	0.952	0.016	0	0	0.016
Group 3	0.036	0.046	0.897	0	0	0.021
Group 4	0.014	0.028	0.003	0.941	0.003	0.011
Group 5	0.01	0.029	0	0	0.941	0.02
Group 6	0	0.162	0.189	0	0	0.649

Accuracy: 94.7%
Precision: 0.950
Recall: 0.947
F1-Score: 0.945

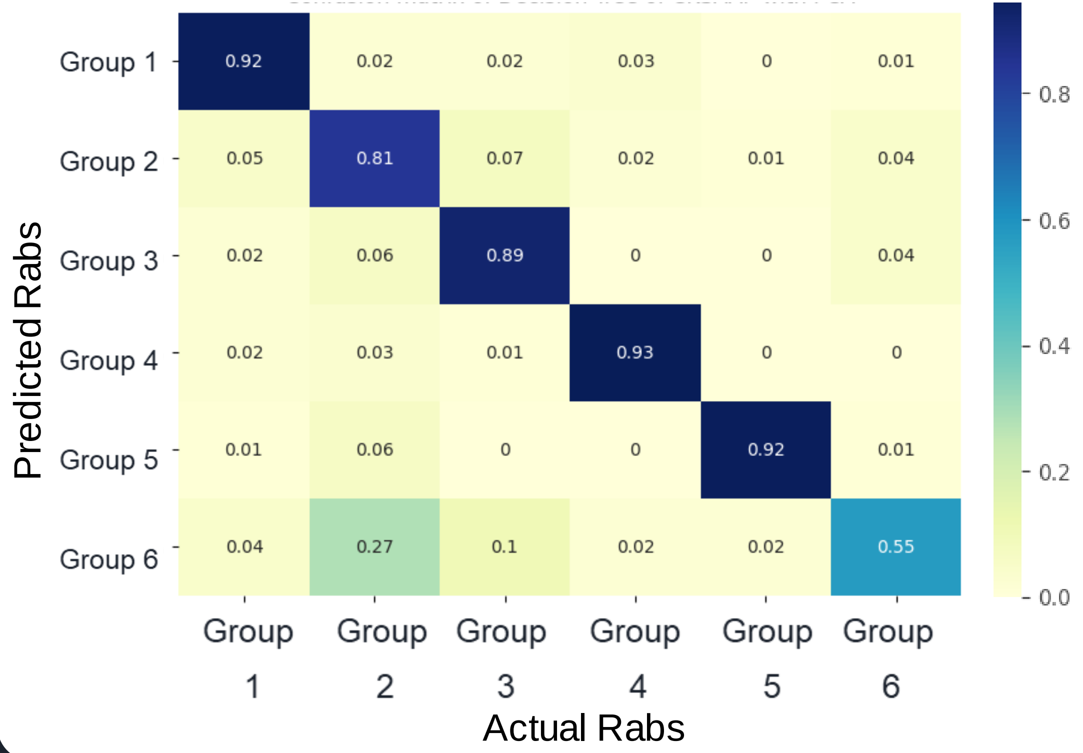
PCA

Accuracy: 93.4%
Precision: 0.936
Recall: 0.934
F1-Score: 0.934

CKSAAP predictor using decision tree



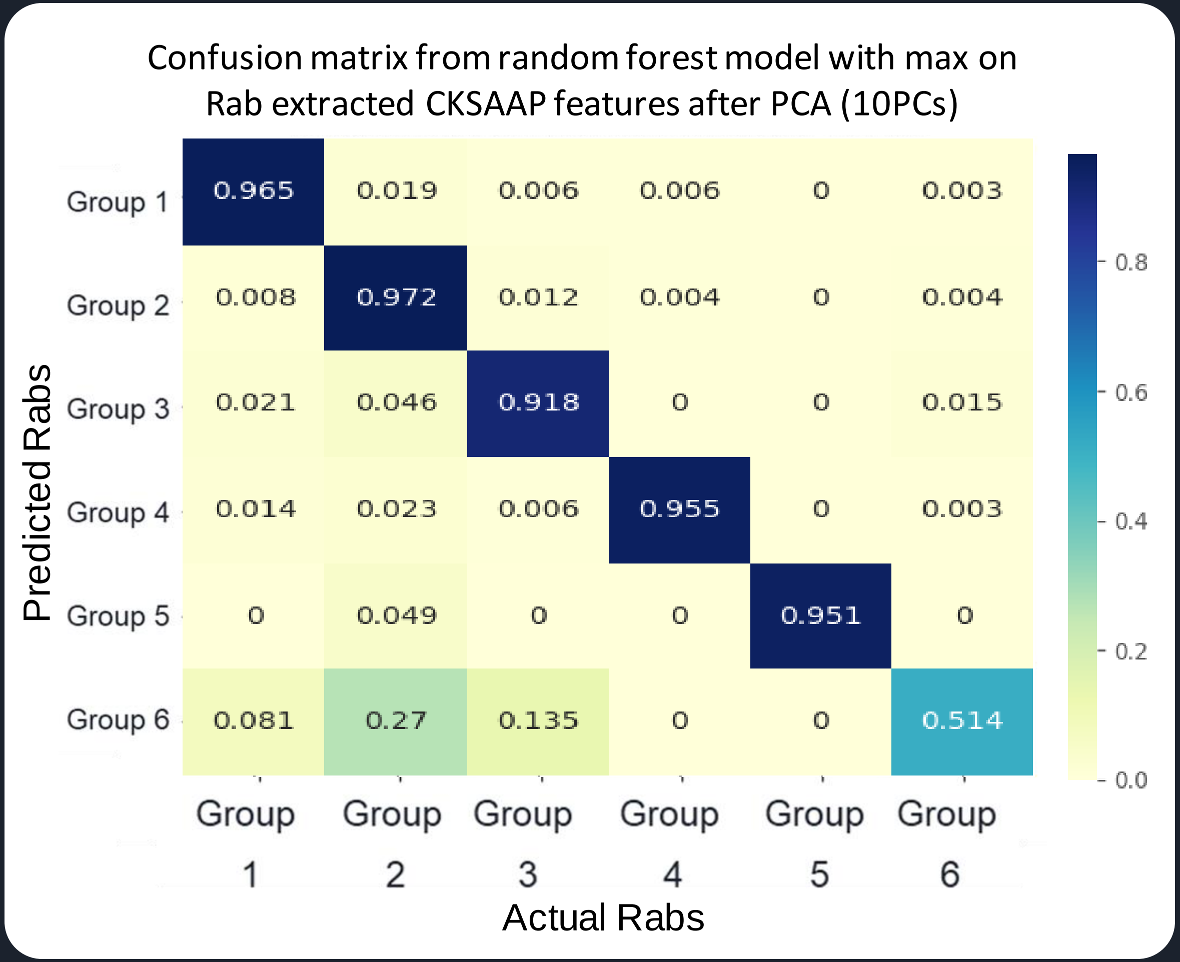
Confusion matrix from decision tree model with max depth = 22 on Rab extracted CKSAAP features after PCA (10PCs)



Predicted Rabs \ Actual Rabs	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
Group 1	0.92	0.02	0.02	0.03	0	0.01
Group 2	0.05	0.81	0.07	0.02	0.01	0.04
Group 3	0.02	0.06	0.89	0	0	0.04
Group 4	0.02	0.03	0.01	0.93	0	0
Group 5	0.01	0.06	0	0	0.92	0.01
Group 6	0.04	0.27	0.1	0.02	0.02	0.55

Accuracy: 89,28%
Precision: 0.839
Recall: 0.835
F-score: 0.837

CKSAAP predictor using random forest



Confusion matrix from random forest model with max on Rab extracted CKSAAP features after PCA (10PCs)

	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
Group 1	0.965	0.019	0.006	0.006	0	0.003
Group 2	0.008	0.972	0.012	0.004	0	0.004
Group 3	0.021	0.046	0.918	0	0	0.015
Group 4	0.014	0.023	0.006	0.955	0	0.003
Group 5	0	0.049	0	0	0.951	0
Group 6	0.081	0.27	0.135	0	0	0.514

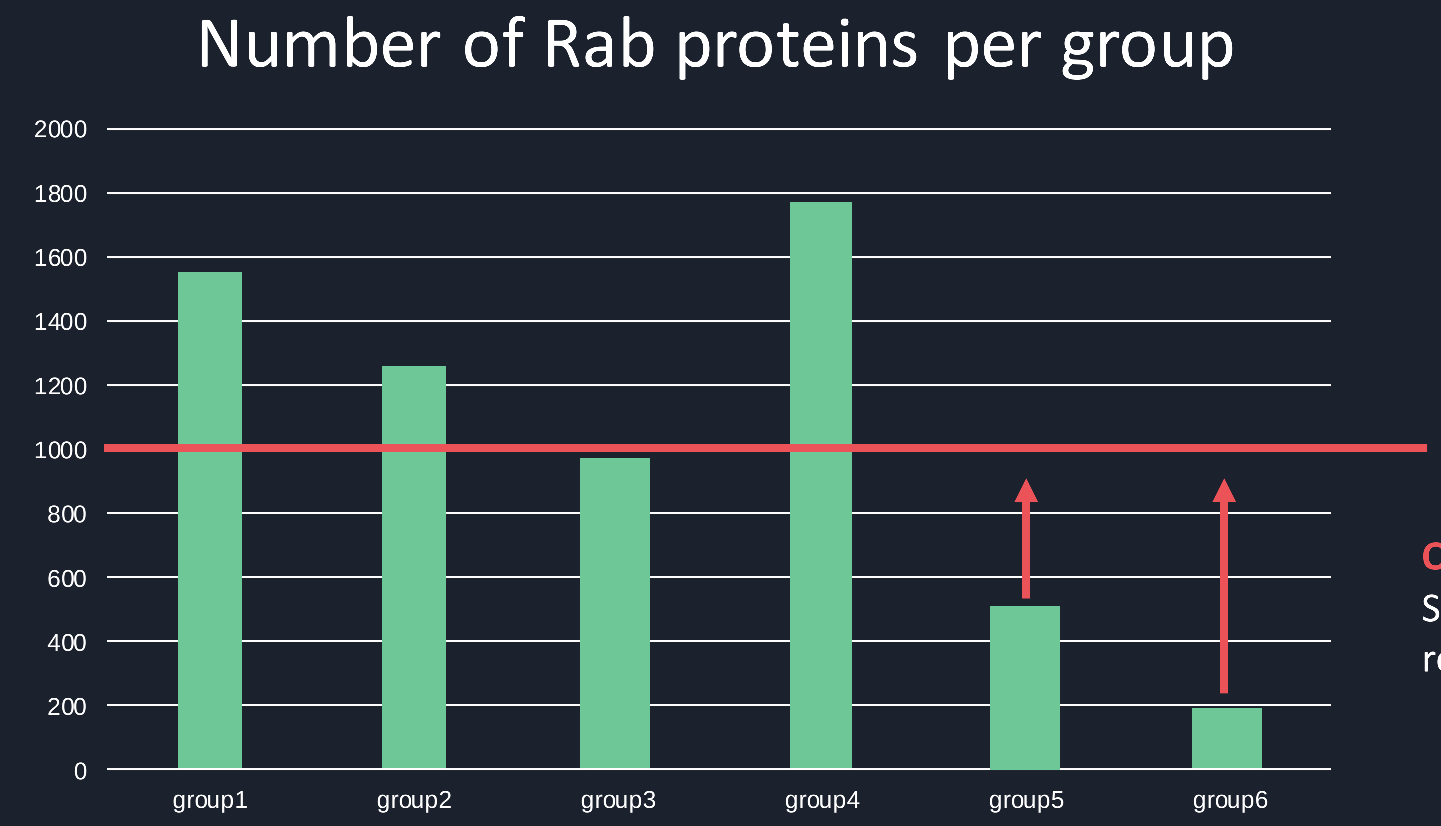
Accuracy : 94.2%

Precision : 0.942

Recall : 0.941

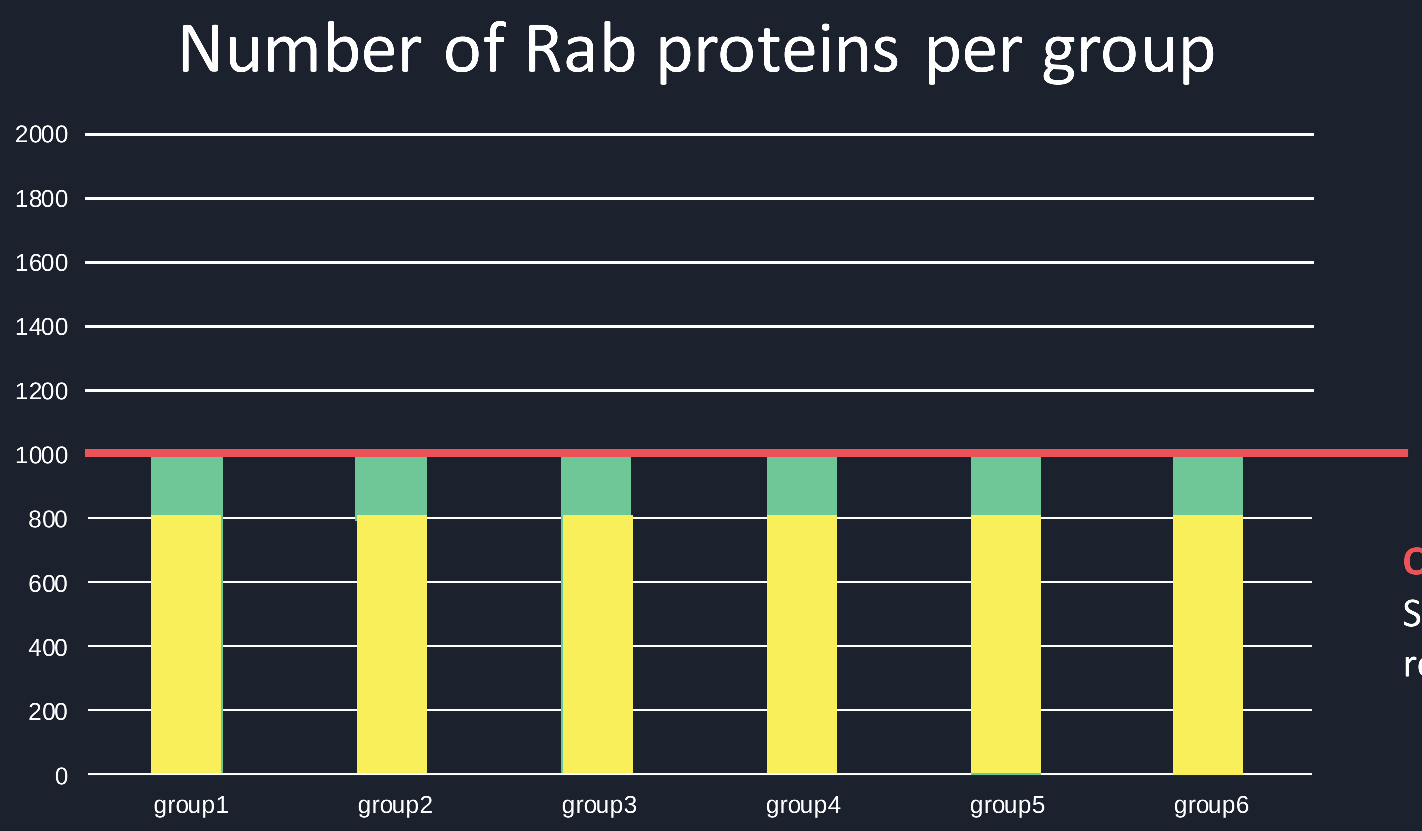
F-score : 0.941

What is going on with group 6?



Oversampling:
Sample with
replacement

What is going on with group 6?

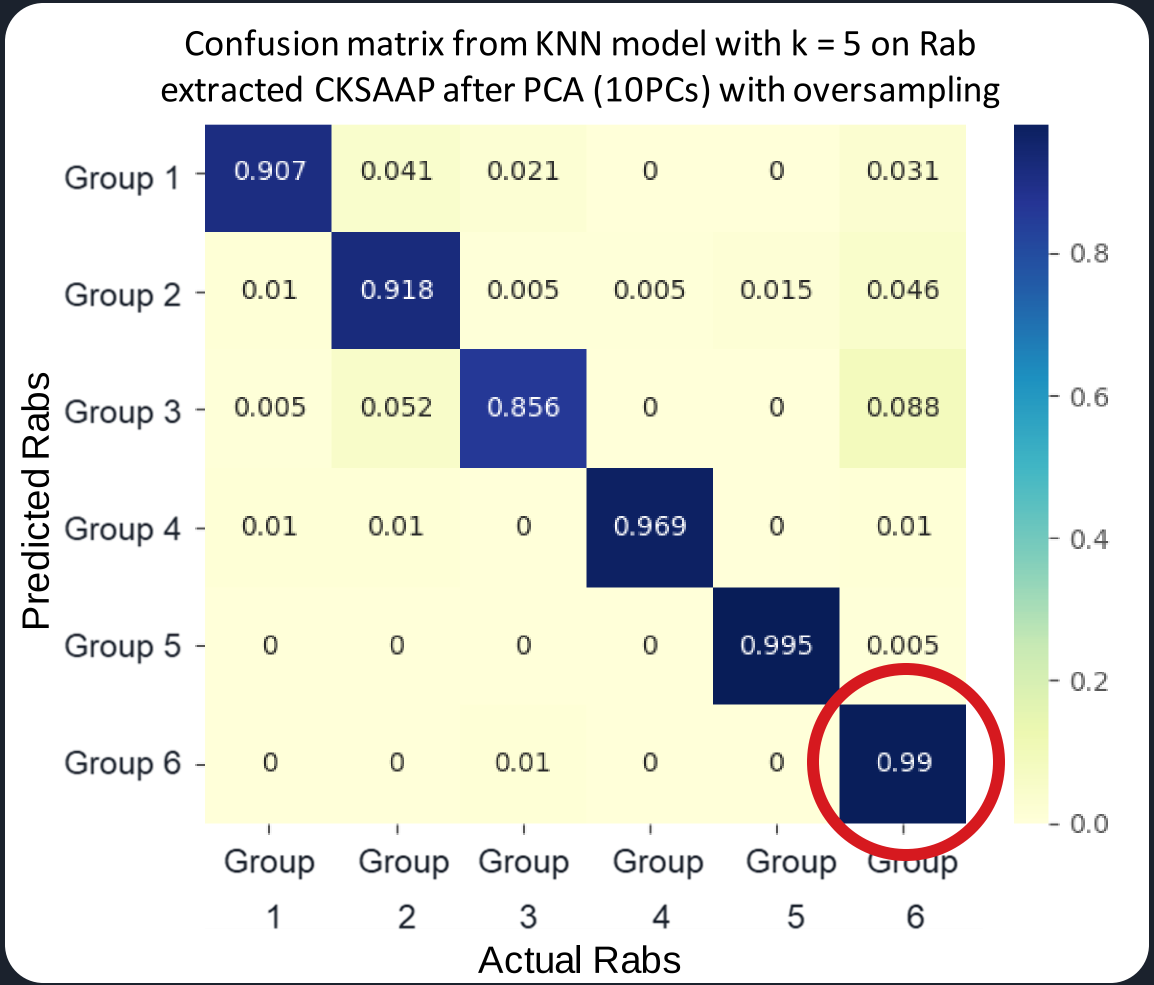


Testing set

Training set

Oversampling:
Sample with
replacement

CKSAAP with ~~resampling~~ overfitting

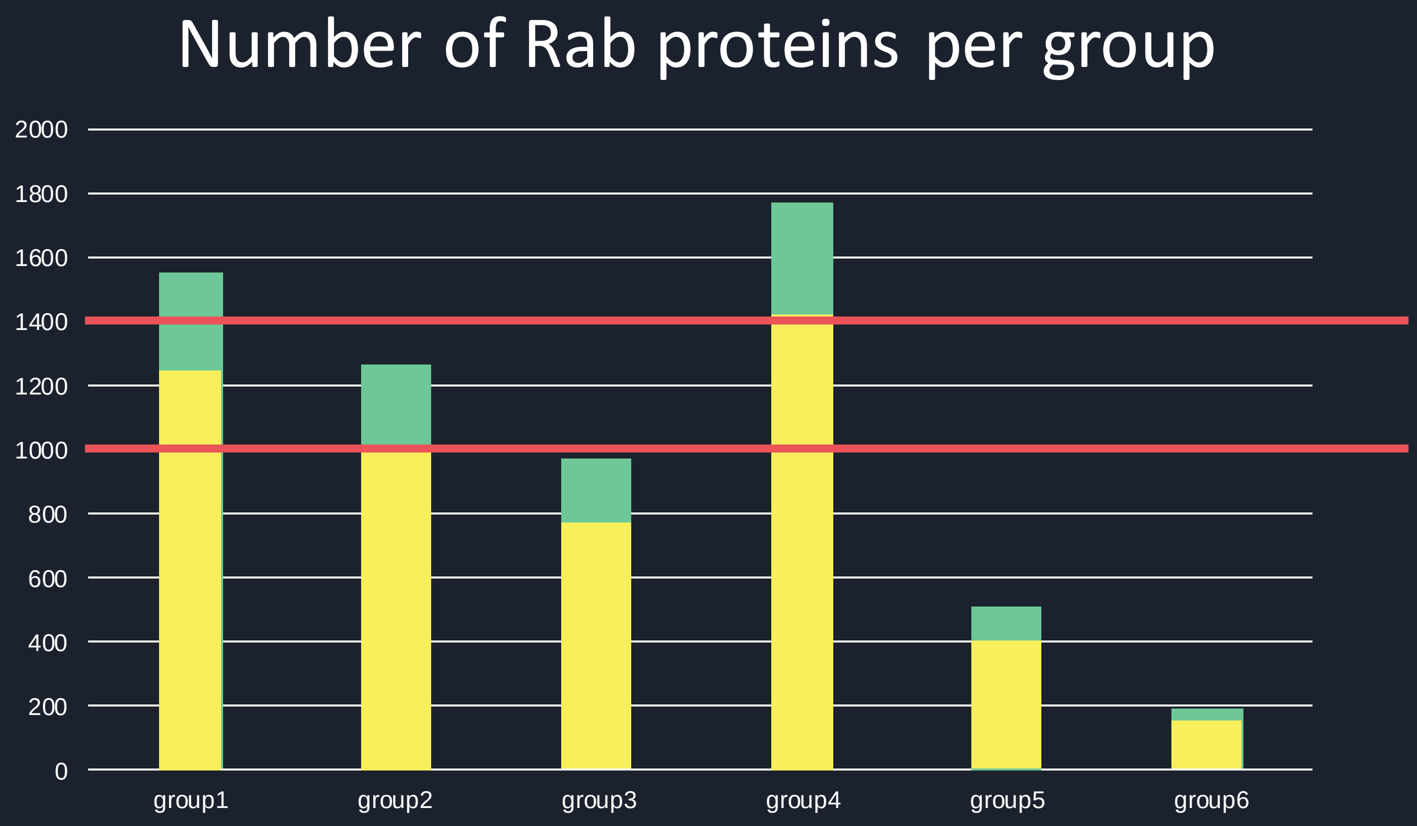


Confusion matrix from KNN model with k = 5 on Rab extracted CKSAAP after PCA (10PCs) with oversampling

Predicted Rabs \ Actual Rabs	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
Group 1	0.907	0.041	0.021	0	0	0.031
Group 2	0.01	0.918	0.005	0.005	0.015	0.046
Group 3	0.005	0.052	0.856	0	0	0.088
Group 4	0.01	0.01	0	0.969	0	0.01
Group 5	0	0	0	0	0.995	0.005
Group 6	0	0	0.01	0	0	0.99

isabeljimenez.com
GIFDB.com

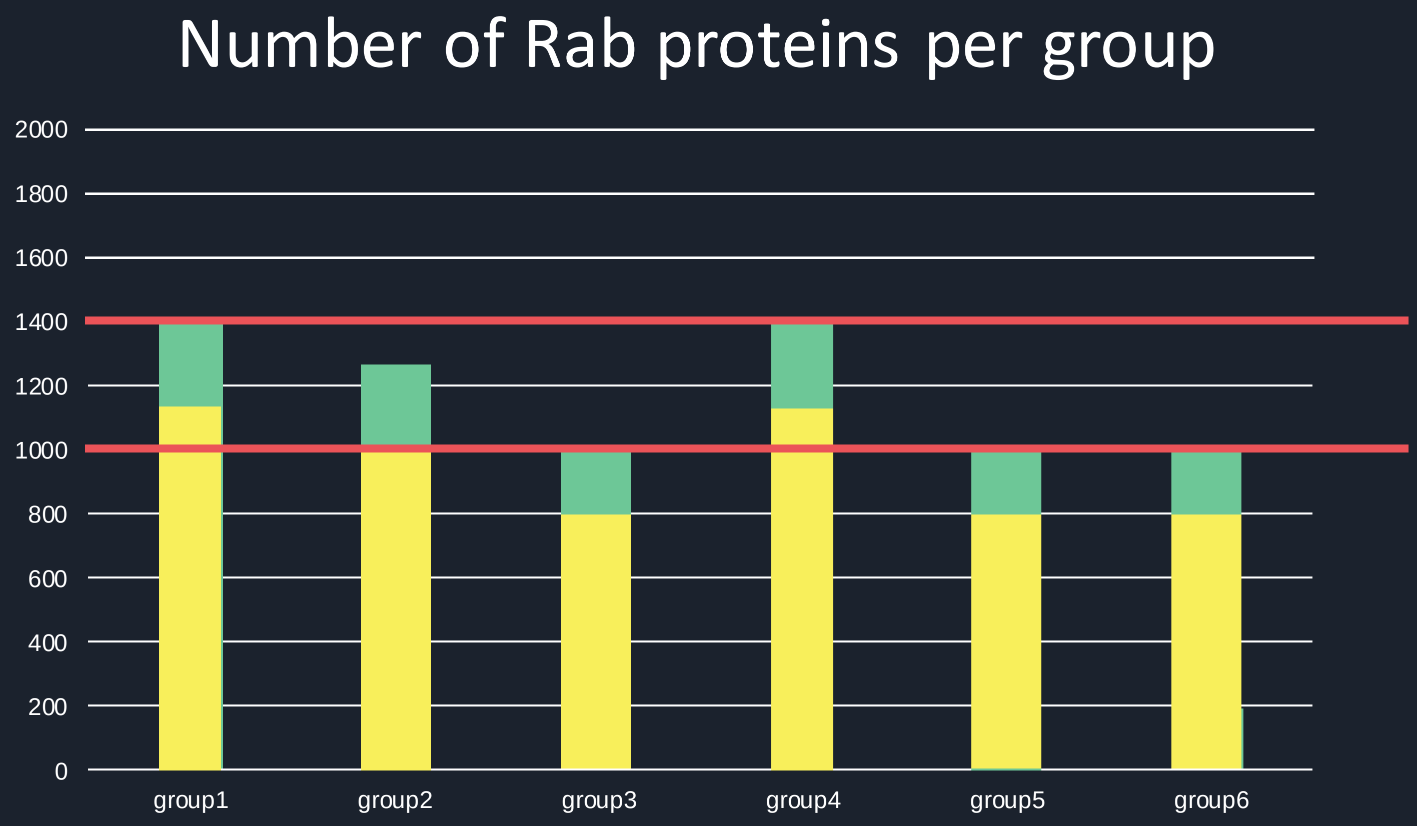
Improving sampling



Testing set

Training set

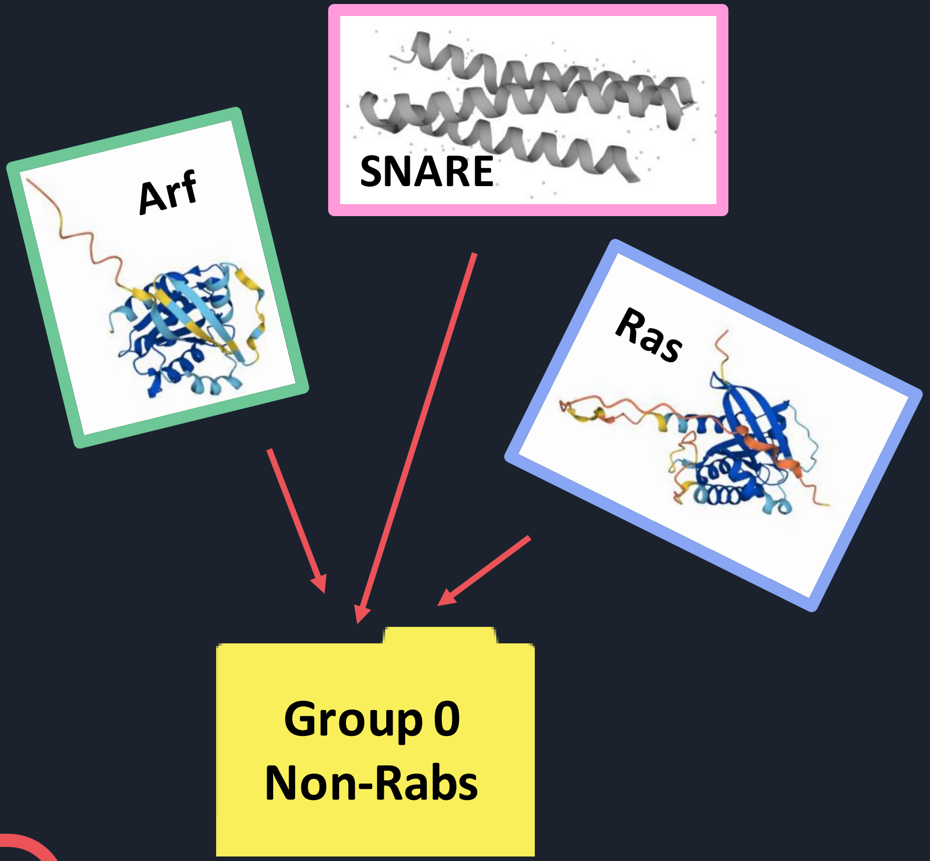
Improving sampling

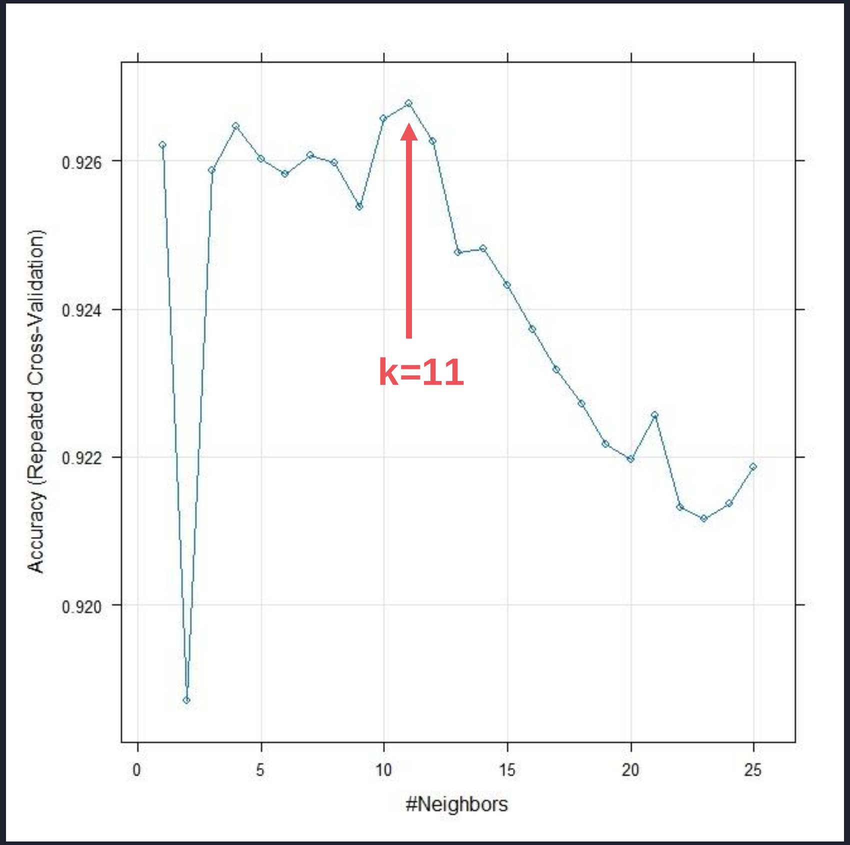


Testing set

Training set

Improving the model

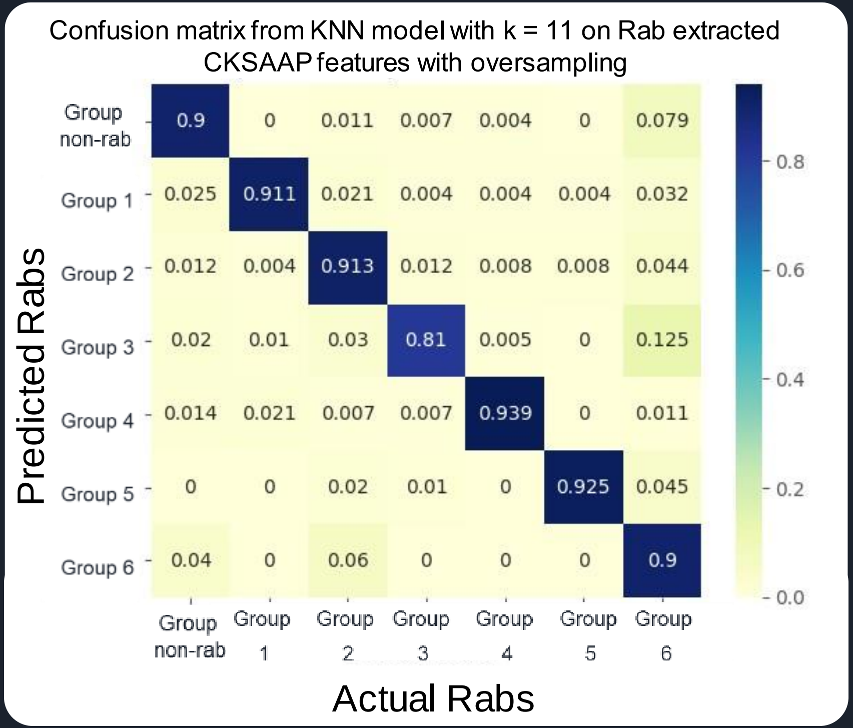




Uniprot
Michel Cuendet

CKSAAP predictor with improved model

KNN



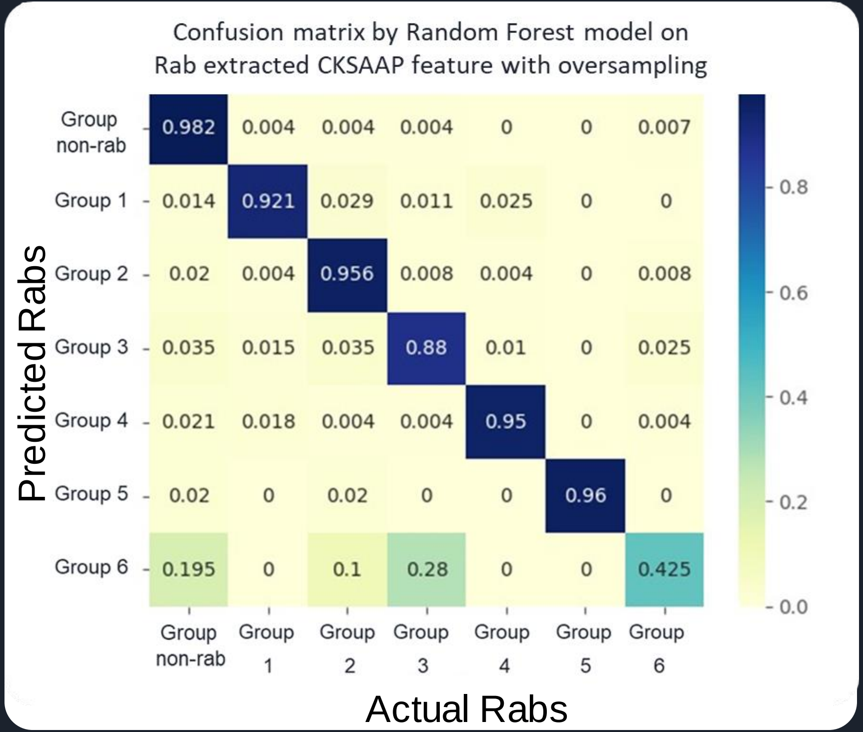
Confusion matrix from KNN model with k = 11 on Rab extracted CKSAAP features with oversampling

	Group non-rab	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
Group non-rab	0.9	0	0.011	0.007	0.004	0	0.079
Group 1	0.025	0.911	0.021	0.004	0.004	0.004	0.032
Group 2	0.012	0.004	0.913	0.012	0.008	0.008	0.044
Group 3	0.02	0.01	0.03	0.81	0.005	0	0.125
Group 4	0.014	0.021	0.007	0.007	0.939	0	0.011
Group 5	0	0	0.02	0.01	0	0.925	0.045
Group 6	0.04	0	0.06	0	0	0	0.9

Accuracy: 90.2% Recall: 0.902

Precision: 0.912 F1-Score: 0.905

Random forest



Confusion matrix by Random Forest model on Rab extracted CKSAAP feature with oversampling

	Group non-rab	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
Group non-rab	0.982	0.004	0.004	0.004	0	0	0.007
Group 1	0.014	0.921	0.029	0.011	0.025	0	0
Group 2	0.02	0.004	0.956	0.008	0.004	0	0.008
Group 3	0.035	0.015	0.035	0.88	0.01	0	0.025
Group 4	0.021	0.018	0.004	0.004	0.95	0	0.004
Group 5	0.02	0	0.02	0	0	0.96	0
Group 6	0.195	0	0.1	0.28	0	0	0.425

Accuracy: 88.2% Recall: 0.882

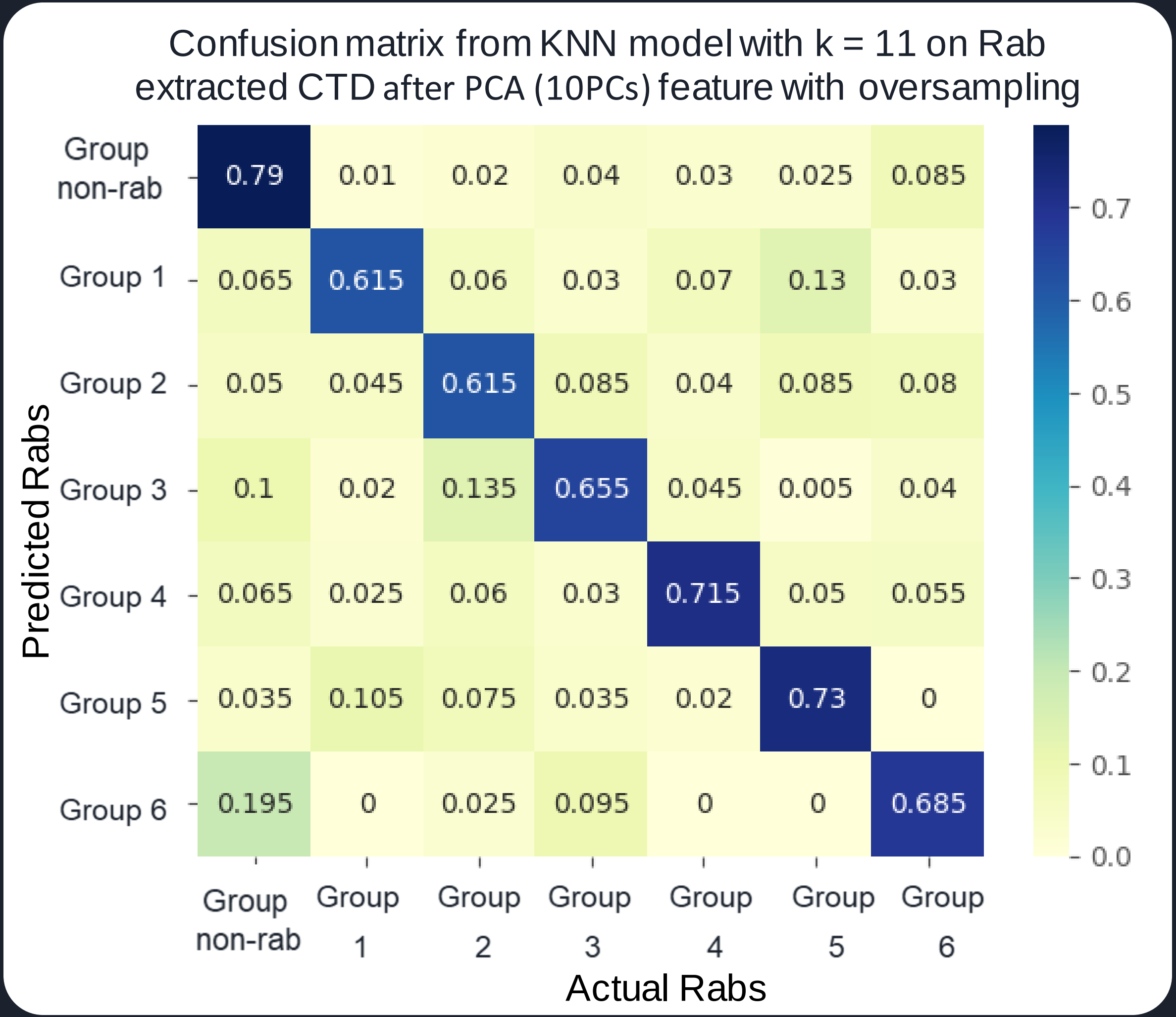
Precision: 0.891 F1-Score: 0.874

Third feature: CTD

- **Physico-chemical** properties (CTD= Conjoint Triad Descriptors)

	Hydro-phobicity	Polarity	Van der Waals volume	Charge	…
rab1	0,212	0,66454	0,117	0,5888	
rab2	0,42588	0,1588	0,3665	0,05	
rab3	0,61	0,258	0,557	0,255	

CTD with KNN



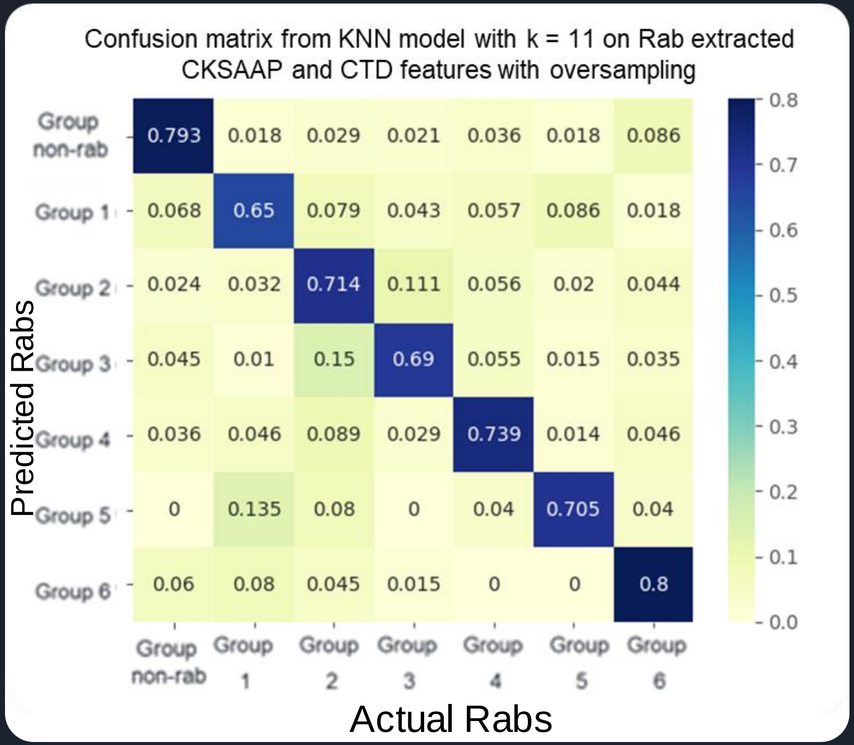
Confusion matrix from KNN model with k = 11 on Rab extracted CTD after PCA (10PCs) feature with oversampling

Predicted Rabs \ Actual Rabs	Group non-rab	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
Group non-rab	0.79	0.01	0.02	0.04	0.03	0.025	0.085
Group 1	0.065	0.615	0.06	0.03	0.07	0.13	0.03
Group 2	0.05	0.045	0.615	0.085	0.04	0.085	0.08
Group 3	0.1	0.02	0.135	0.655	0.045	0.005	0.04
Group 4	0.065	0.025	0.06	0.03	0.715	0.05	0.055
Group 5	0.035	0.105	0.075	0.035	0.02	0.73	0
Group 6	0.195	0	0.025	0.095	0	0	0.685

Accuracy: 69%
Precision: 0.692
Recall: 0.686
F1-Score: 0.686

CKSAAP and CTD predictors combined

KNN



Confusion matrix from KNN model with k = 11 on Rab extracted CKSAAP and CTD features with oversampling

	Group non-rab	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
Group non-rab	0.793	0.018	0.029	0.021	0.036	0.018	0.086
Group 1	0.068	0.65	0.079	0.043	0.057	0.086	0.018
Group 2	0.024	0.032	0.714	0.111	0.056	0.02	0.044
Group 3	0.045	0.01	0.15	0.69	0.055	0.015	0.035
Group 4	0.036	0.046	0.089	0.029	0.739	0.014	0.046
Group 5	0	0.135	0.08	0	0.04	0.705	0.04
Group 6	0.06	0.08	0.045	0.015	0	0	0.8

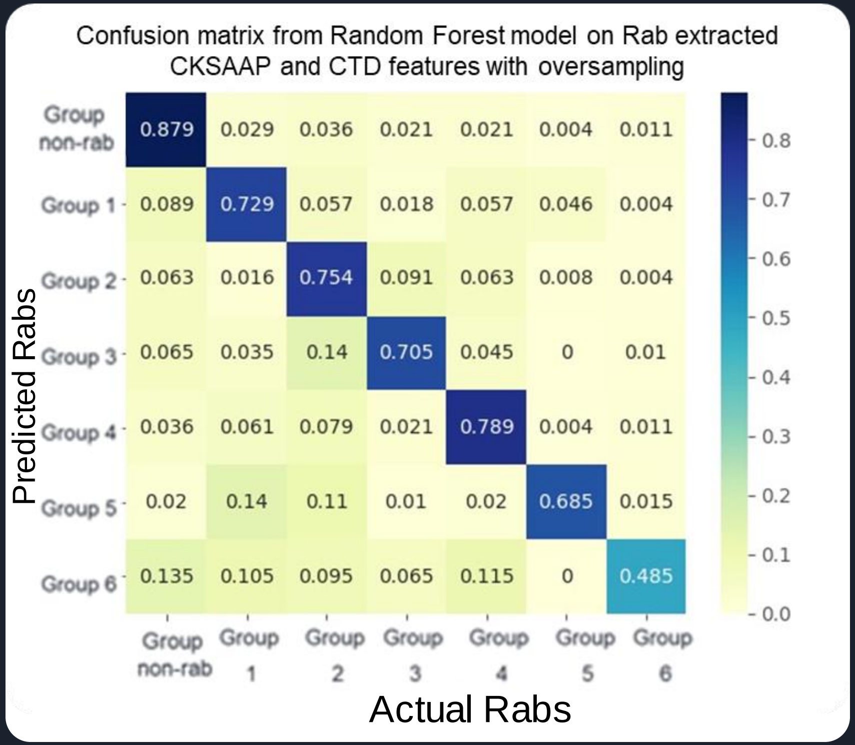
Accuracy: 73%

Precision: 0.731

Recall: 0.727

F1-Score: 0.727

Random forest



Confusion matrix from Random Forest model on Rab extracted CKSAAP and CTD features with oversampling

	Group non-rab	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
Group non-rab	0.879	0.029	0.036	0.021	0.021	0.004	0.011
Group 1	0.089	0.729	0.057	0.018	0.057	0.046	0.004
Group 2	0.063	0.016	0.754	0.091	0.063	0.008	0.004
Group 3	0.065	0.035	0.14	0.705	0.045	0	0.01
Group 4	0.036	0.061	0.079	0.021	0.789	0.004	0.011
Group 5	0.02	0.14	0.11	0.01	0.02	0.685	0.015
Group 6	0.135	0.105	0.095	0.065	0.115	0	0.485

Accuracy: 73%

Precision: 0.747

Recall: 0.730

F1-Score: 0.728

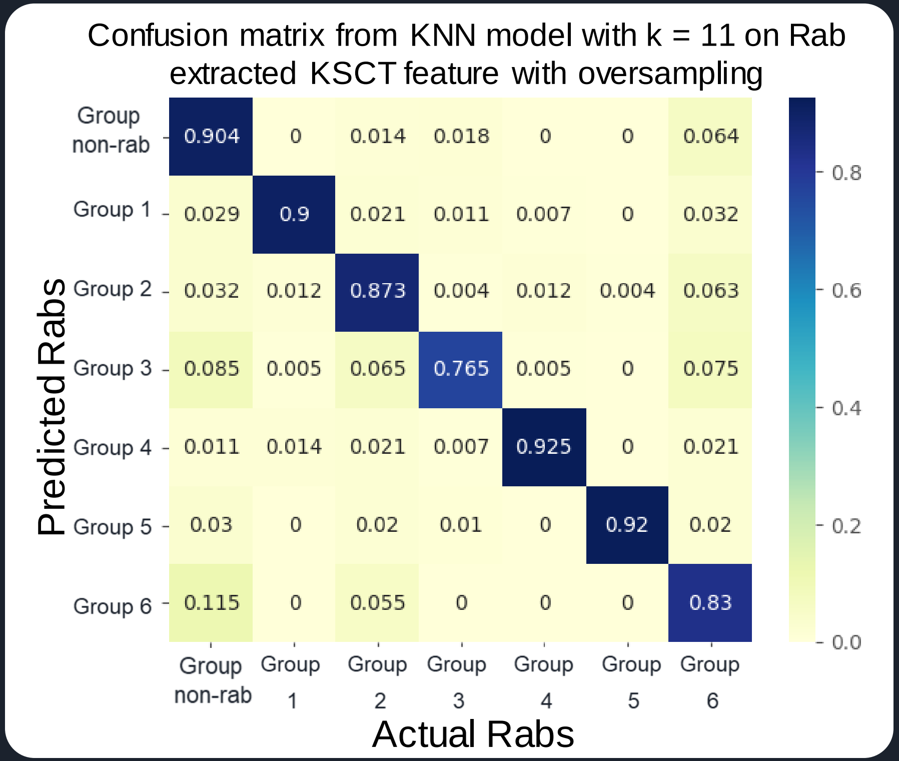
Fourth feature: KSCT

- **Physico-chemical** properties (KSCT= K spaced conjoint triad)

	Hydro-phobicity	Polarity	Van der Waals volume	Charge	…
rab1	0,212	0,66454	0,11	0,5888	
rab2	0,02588	0,0588	0,3665	0,005	
rab3	0,6	0,258	0,557	0,255	

KSCT predictor

KNN


Confusion matrix from KNN model with k = 11 on Rab extracted KSCT feature with oversampling

Predicted Rabs \ Actual Rabs	Group non-rab	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
Group non-rab	0.904	0	0.014	0.018	0	0	0.064
Group 1	0.029	0.9	0.021	0.011	0.007	0	0.032
Group 2	0.032	0.012	0.873	0.004	0.012	0.004	0.063
Group 3	0.085	0.005	0.065	0.765	0.005	0	0.075
Group 4	0.011	0.014	0.021	0.007	0.925	0	0.021
Group 5	0.03	0	0.02	0.01	0	0.92	0.02
Group 6	0.115	0	0.055	0	0	0	0.83

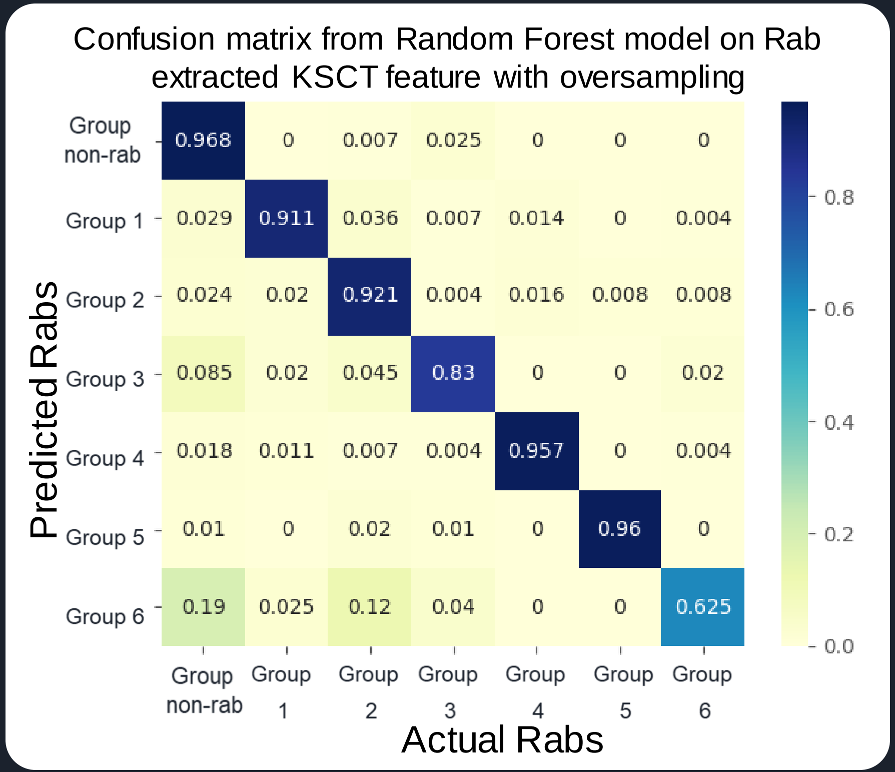
Accuracy: 87.9%

Recall: 0.879

Precision: 0.888

F1-Score: 0.881

Random forest


Confusion matrix from Random Forest model on Rab extracted KSCT feature with oversampling

Predicted Rabs \ Actual Rabs	Group non-rab	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
Group non-rab	0.968	0	0.007	0.025	0	0	0
Group 1	0.029	0.911	0.036	0.007	0.014	0	0.004
Group 2	0.024	0.02	0.921	0.004	0.016	0.008	0.008
Group 3	0.085	0.02	0.045	0.83	0	0	0.02
Group 4	0.018	0.011	0.007	0.004	0.957	0	0.004
Group 5	0.01	0	0.02	0.01	0	0.96	0
Group 6	0.19	0.025	0.12	0.04	0	0	0.625

Accuracy: 89.2%

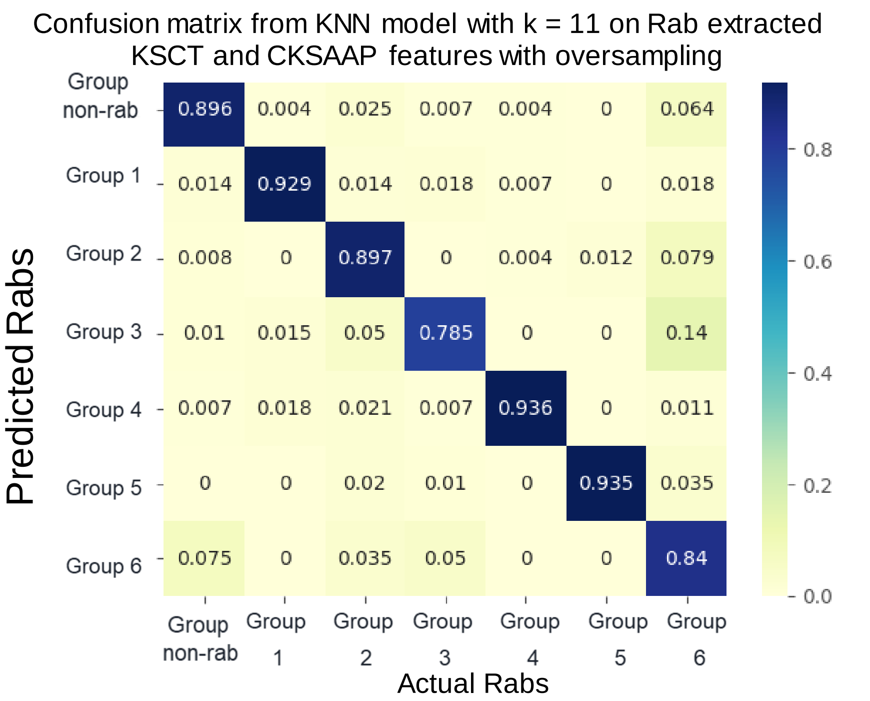
Recall: 0.892

Precision: 0.900

F1-Score: 0.890

KSCT and CKSAAP predictors combined

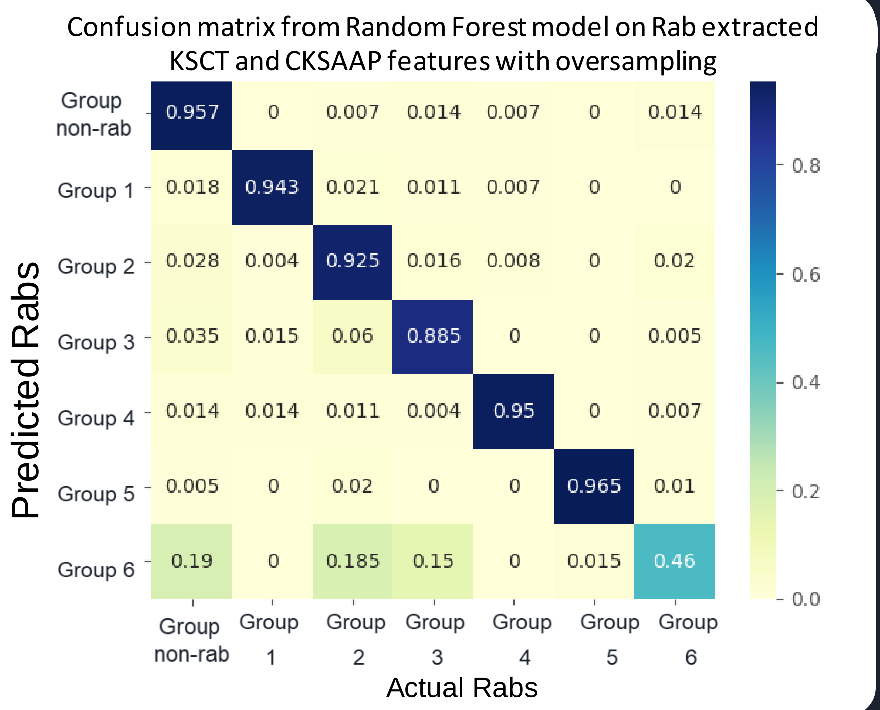
KNN


Confusion matrix from KNN model with k = 11 on Rab extracted KSCT and CKSAAP features with oversampling

Predicted Rabs \ Actual Rabs	Group non-rab	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
Group non-rab	0.896	0.004	0.025	0.007	0.004	0	0.064
Group 1	0.014	0.929	0.014	0.018	0.007	0	0.018
Group 2	0.008	0	0.897	0	0.004	0.012	0.079
Group 3	0.01	0.015	0.05	0.785	0	0	0.14
Group 4	0.007	0.018	0.021	0.007	0.936	0	0.011
Group 5	0	0	0.02	0.01	0	0.935	0.035
Group 6	0.075	0	0.035	0.05	0	0	0.84

Accuracy: 89.3%
Precision: 0.901
Recall: 0.893
F1-Score: 0.895

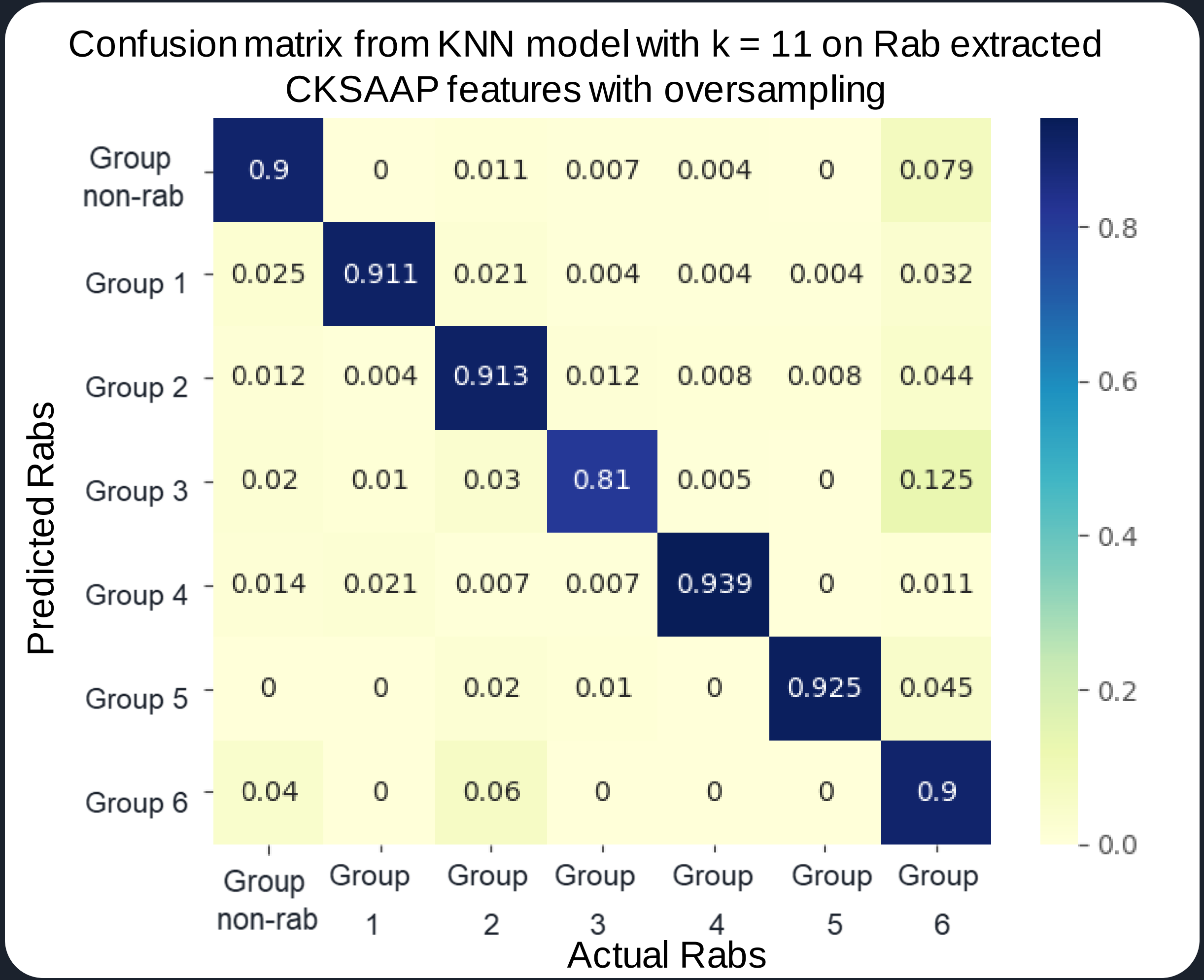
Random forest


Confusion matrix from Random Forest model on Rab extracted KSCT and CKSAAP features with oversampling

Predicted Rabs \ Actual Rabs	Group non-rab	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
Group non-rab	0.957	0	0.007	0.014	0.007	0	0.014
Group 1	0.018	0.943	0.021	0.011	0.007	0	0
Group 2	0.028	0.004	0.925	0.016	0.008	0	0.02
Group 3	0.035	0.015	0.06	0.885	0	0	0.005
Group 4	0.014	0.014	0.011	0.004	0.95	0	0.007
Group 5	0.005	0	0.02	0	0	0.965	0.01
Group 6	0.19	0	0.185	0.15	0	0.015	0.46

Accuracy: 88.2%
Precision: 0.888
Recall: 0.882
F1-Score: 0.876

Best of the best: CKSAAP with oversampling



Confusion matrix from KNN model with k = 11 on Rab extracted CKSAAP features with oversampling

Predicted Rabs \ Actual Rabs	Group non-rab	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
Group non-rab	0.9	0	0.011	0.007	0.004	0	0.079
Group 1	0.025	0.911	0.021	0.004	0.004	0.004	0.032
Group 2	0.012	0.004	0.913	0.012	0.008	0.008	0.044
Group 3	0.02	0.01	0.03	0.81	0.005	0	0.125
Group 4	0.014	0.021	0.007	0.007	0.939	0	0.011
Group 5	0	0	0.02	0.01	0	0.925	0.045
Group 6	0.04	0	0.06	0	0	0	0.9

Accuracy: 90.2%

Precision: 0.912

Recall: 0.902

F1-Score: 0.905

Take home messages

- Best model: KNN with CKSAAP feature
- Combos of features do not add to the model

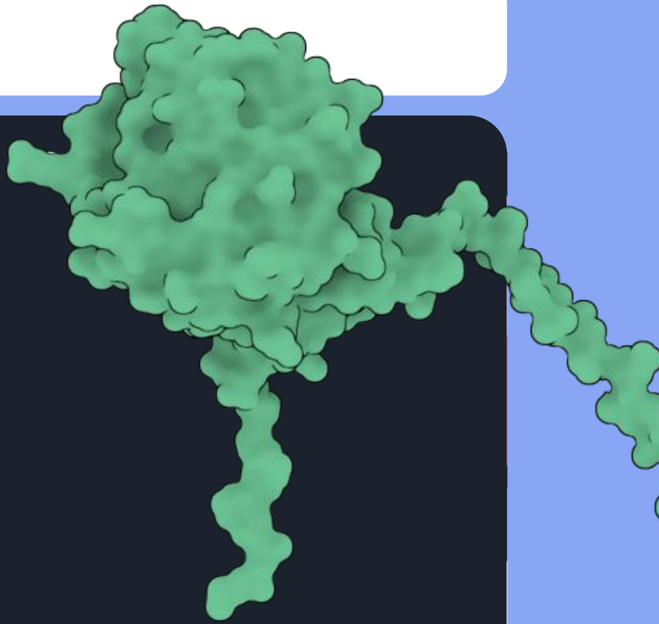
Challenges

- **Unequal distribution** in classes
- **Unintuitive results**: best model is not a combination (?)
- **Lots of testing** and probing, limited time
- **Google colab**… not so easy to colab
- **Different kind of coding** but lots of help from our mentor and forums/Chat GPT

Feedback

- **Highly concentrated working period** blur lines between work and personal time
- **Grade disproportionally based on presentation,** rather than work done
- **Little time** between intermediate and final presentation but **helpful and inspiring**

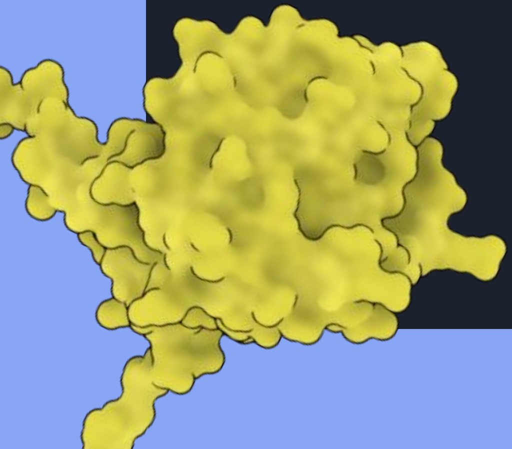
- Very **helpful mentor**
- **Apply theory** seen in class on an interesting project
- Learned a lot about **coding and problem solving**



Thanks!

Questions?

CREDITS: This presentation template was created by **Slidesgo,** including icons by **Flaticon,** and infographics & images by **Freepik**



Email: gaetan.ray@unil.ch, luca.paley@unil.ch, sydney.fleming@unil.ch