Reliability of putative RRIs

Final presentation of Masterproject
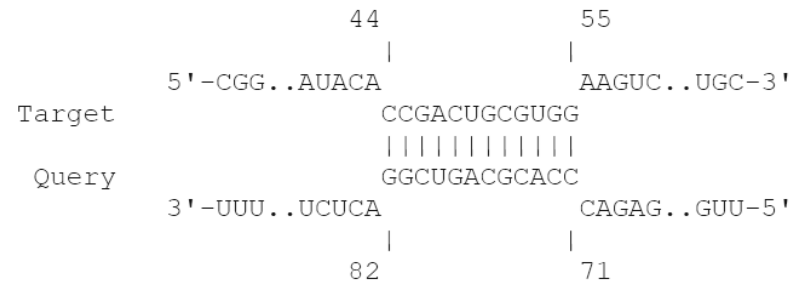29.07.2021
Sebastian Holler
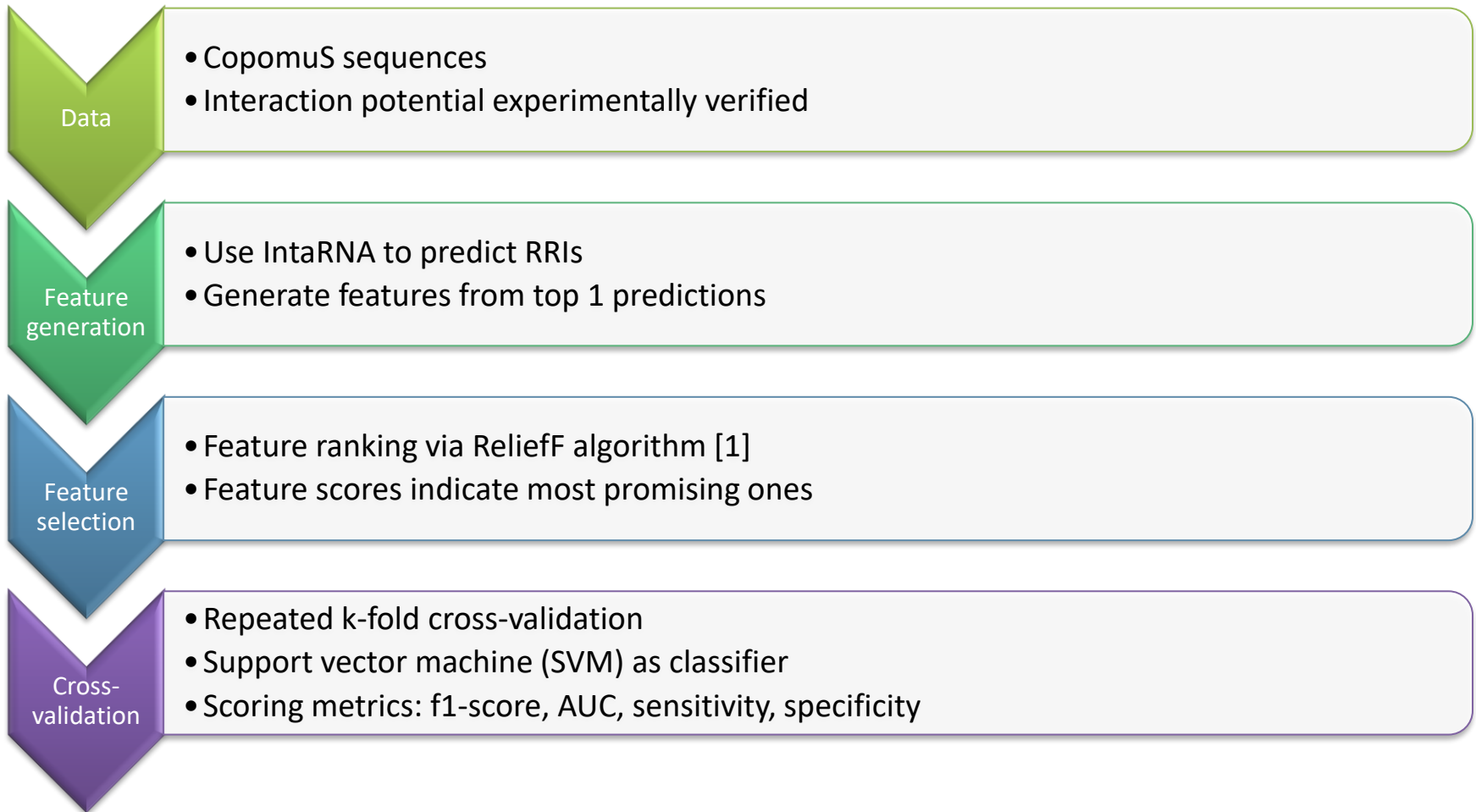
Supervisors: Florian Eggenhofer, Teresa Müller, Martin Raden
Chair of Bioinformatics

Motivation

```
                 44              55
                 |               |
        5'-CGG..AUACA       AAGUC..UGC-3'
Target             CCGACUGCGUGG
                   ||||||||||||
 Query             GGCUGACGCACC
        3'-UUU..UCUCA       CAGAG..GUU-5'
                 |               |
                82              71
```

?

Workflow

Data
- CopomuS sequences
- Interaction potential experimentally verified

Feature generation
- Use IntaRNA to predict RRIs
- Generate features from top 1 predictions

Feature selection
- Feature ranking via ReliefF algorithm [1]
- Feature scores indicate most promising ones

Cross-validation
- Repeated k-fold cross-validation
- Support vector machine (SVM) as classifier
- Scoring metrics: f1-score, AUC, sensitivity, specificity

[1]: Benchmarking relief-based feature selection methods for bioinformatics data mining: doi.org/10.1016/j.jbi.2018.07.015
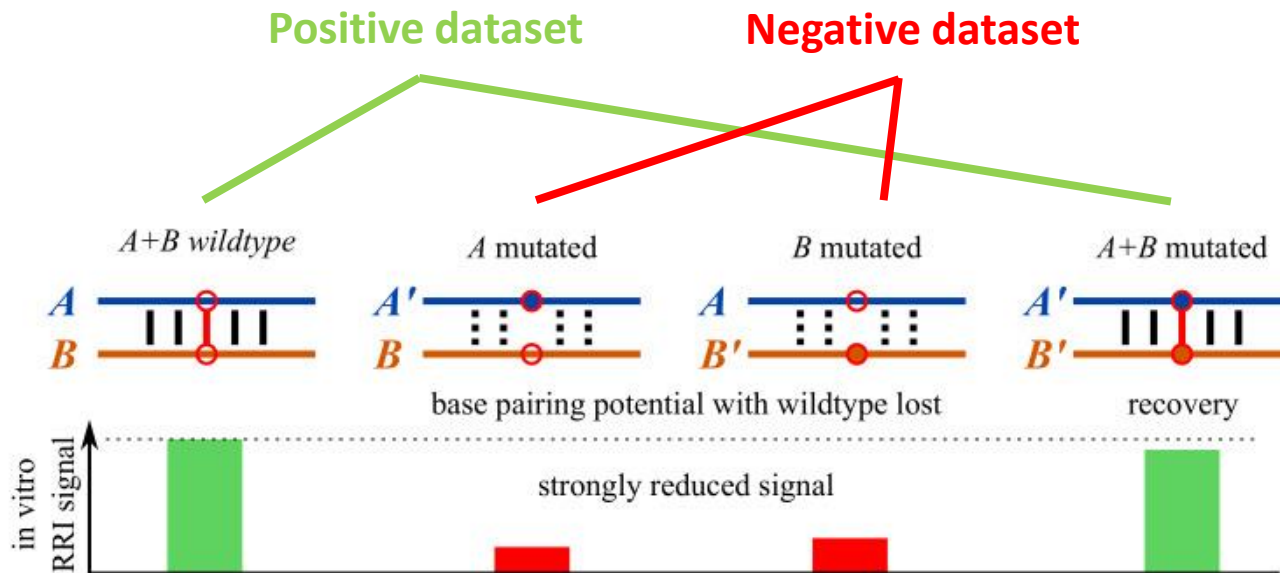
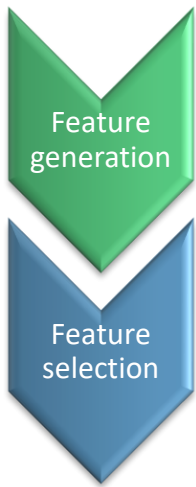# Interaction data

Data

- Challenge: Lack of experimental verified data
- Use data from compensatory mutation experiments

Positive dataset **Negative dataset**

CopomuS-Ranking Compensatory Mutations to Guide RNA-RNA Interaction Verification
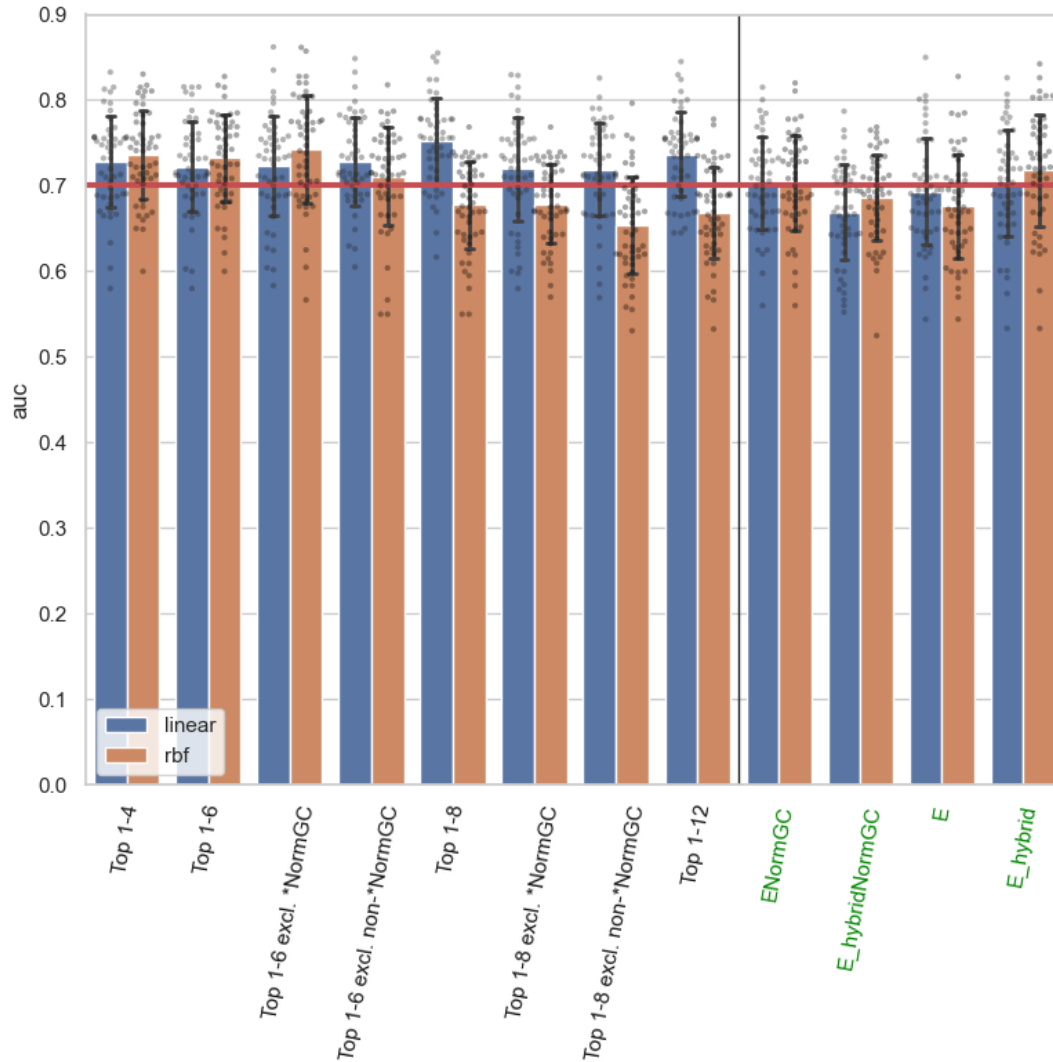Experiments: pubmed.ncbi.nlm.nih.gov/32481751/

ReliefF results

Feature name	Description	Importance
ENormGC	E normalized by interGC	0.12430758
E_hybridNormGC	E_hybrid normalized by interGC	0.1114999
E	Minimum free energy	0.10917744
E_hybrid	Hybridization energy	0.0989977
maxEDNormGC	Max. accessibility energy normalized by interGC	0.04220053
maxED	Max. accessibility energy	0.03550688
numBP	Number of base pairs within interaction	0.01767762
interGC	GC-content within interaction	0.01613493
diffGC	Difference between interGC and globalGC	0.00770325
minRange	Min. length of interacting subsequences	0.006607
maxRange	Max. length of interacting subsequences	0.00314323
minRangeNorm	minRange normalized by numBP	0.00284907
distSCodon	Signed distance start codon - interaction	0.00000325
distSCodonAbs	Absolute distance start codon - interaction	0.00000325
globalGC	GC-content of whole input sequences	-0.00253847
minRangeNormInv	Inverse minRange normalized by numBP	-0.00557687
maxRangeNorm	maxRange normalized by numBP	-0.01102459
maxRangeNormInv	Inverse maxRange normalized by numBP	-0.01213895

- **Energy-based features** top ranking
- **Features normalized by interGC** promising

Cross-validation results

Conclusion

Data
- Data from compensatory mutation experiments
- Interaction potential experimentally verified

Feature generation
- Use IntaRNA to predict RRIs
- Generate features from top 1 predictions

Feature selection
- ReliefF ranking topped by energy features
- Energies normalized by interGC promising

Cross-validation
- Energy alone quite good already
- Hard to get better than baseline
- Top 1-8 feature set performs best