# Graph Neural Networks for Identification of Robust Biomarkers from Multi-Omics Data

Lukáš Růžička

# Datasets

- **5 datasets:**
  - allRNA – 58216 features
  - circRNA – 51018 features
  - miRNA – 2553 features
  - piTNA – 556 features
  - TE – 687 features
- **Data complete for 66 subjects**

# Classification tasks

- **Disease**
  - Binary classification
  - Class imbalance (7/59)
- **Risk, Mutation**
  - # classes = 3

# Baseline models

- **Multinomial logistic regression**
  - L1 regularization
- **Support vector classifier**
  - Recursive feature elimination
- **Gradient boosting classifier**
  - Gini importance (impurity-based feature selection)

# Baseline classification

- **Classification performed on:**
  - Each omic type individually
  - Joint dataset
  - For each task, 5 iterations with 6-k cross-validation
  - Feature selection performed for 4 different settings of selection “strength”
    - LR → C (inverse of lambda in Lasso)
    - SVC, GBC → n (number of highest ranking features)

# Baseline results (joint only)

| | classifier | parameter | precision | recall | F-1 |
|---|---|---|---|---|---|
| Disease | MLR | C = 0.6 | 0.80 +- .00 | 0.84 +- .00 | 0.84 +- .00 |
| | SVC | Number of features = 100 | 0.94 +- .00 | 0.94 +- .00 | 0.93 +- .00 |
| | GBC | Number of features = 20 | 0.98 +- .01 | 0.98 +- .01 | 0.99 +- .01 |
| Risk | MLR | C = 1 | 0.36 +- .00 | 0.45 +- .00 | 0.35 +- .00 |
| | SVC | Number of features = 100 | 0.91 +- .00 | 0.89 +- .00 | 0.89 +- .00 |
| | GBC | Number of features = 10 | 0.86 +- .02 | 0.84 +- .02 | 0.84 +- .02 |
| Mutation | MLR | C = 0.8 | 0.39 +- .00 | 0.62 +- .00 | 0.48 +- .00 |
| | SVC | Number of features = 100 | 0.86 +- .00 | 0.82 +- .00 | 0.80 +- .00 |
| | GBC | Number of features = 100 | 0.91 +- .03 | 0.90 +- .02 | 0.90 +- .03 |

# MOGONET performance

- Tested on all datasets together
- MRMR used for feature selection (top 200 features)
- # of training epochs: 1000
- # of pretrain epochs: 500
- Parameter k set to 2, 3, 5 and 10

# MOGONET performace results (best)

| | k | Accuracy | F1 | AUC | F1 weighted | F1 macro |
|---|---|---|---|---|---|---|
| Disease | 10 | 0.94 | 0.97 | 1.00 | - | - |
| Risk | 2 | 0.94 | - | - | 0.94 | 0.95 |
| Mutation | 2 | 0.90 | - | - | 0.89 | 0.89 |

+ MOGONET was able to rank 30 most important biomarkers

# Other frameworks that look promising

- **GCNCC**
  - + Works with PPI
  - - only binary classification
  - - input not specified
- **ScGNN**
  - + intuitive interface
  - + example input data look similar to ours
  - + output seems suitable
    - Graph, learned embeddings, identified cell types

# Other frameworks that look promising

- **PAMOGK**
  - - MOSEK Optimizer API – should be free for students upon request
  - - Input data format not suitable
  - - Don’t have pathway data file (ndexbio.org)
- **DeepOmix**
  - - functional modules (prior knowledge) must be provided by user
    - (tissue network, gene co-expression network, signaling pathways
  - + good UI
  - + also outputs top-ranked identified pathways

# Other frameworks that look promising

- **GCN_Cancer**
  - + Utilizes PPI – however, needs to be generated by user
    - Partial description of how to do it is given
  - Only some datasets could be probably used (ensemble gene ID)

# Further goals

- **Gather results from more GNN frameworks**
  - ScGNN, DeepOmix and GCN_cancer frameworks look the most promising
- **Validate if the marked biomarkers coincide across all the frameworks**
- **Discuss their validity with expert**
- **Possibly try out different combinations of omic types used in analysis**