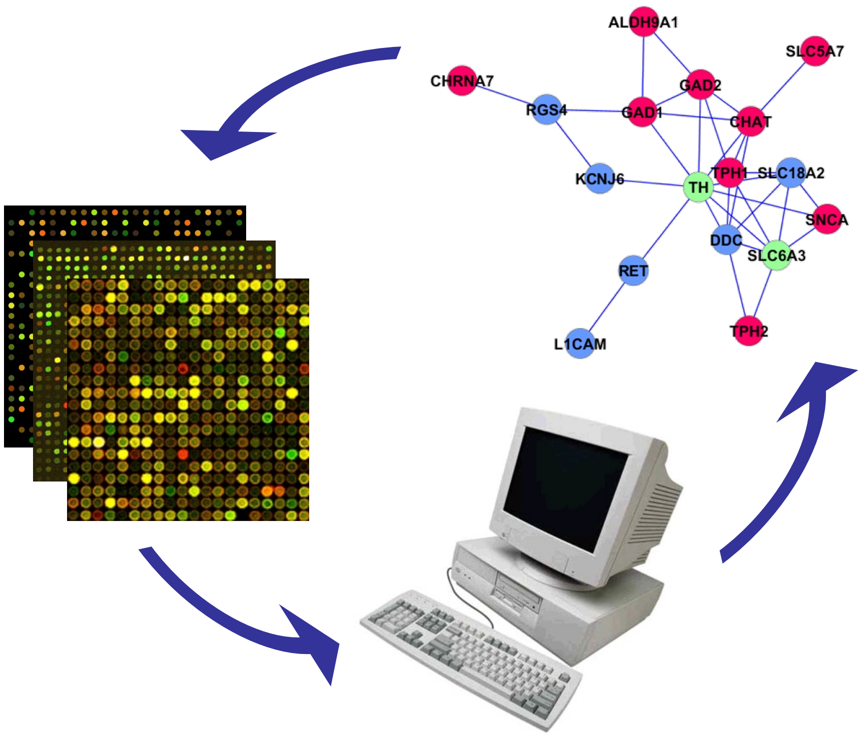


Omics network analysis using mathematical programming

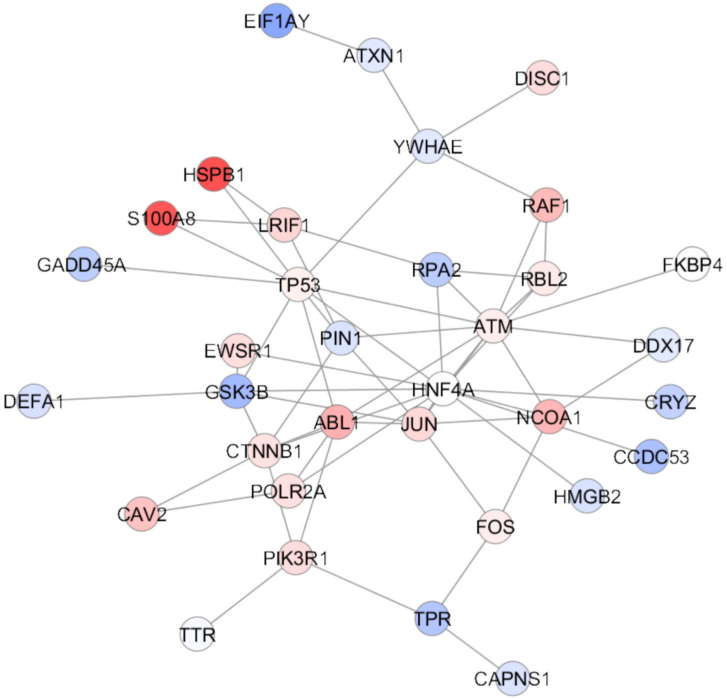
Nikos Vlassis & Enrico Glaab, Luxembourg Centre for Systems Biomedicine

UNIVERSITÉ DU LUXEMBOURG

LCSB

Motivation

Disease-associated molecular perturbations are often localized in biological networks. Finding these network clusters may help us to develop more robust biomarker models.



Example sub-network for Parkinson‘s disease (PD):

🔴 Over-expressed in PD

🔵 Under-expressed in PD

Question: How can we find clustered gene/protein groups efficiently, accounting for their predictivity and connectedness in the network?

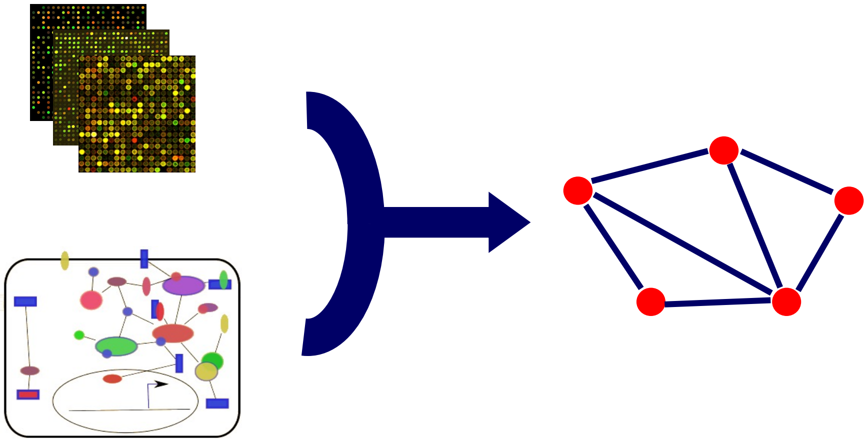
GenePEN - Workflow

Input:

- Gene/protein expression dataset **X** (p rows = genes, n columns = samples)
- Class labels **y** (e.g., “patient vs. control”, “disease subtype 1 vs. disease subtype 2”)
- Table **A** of interactions/similarities between rows in X (e.g., protein-protein interactions)

Output:

- A subset of discriminative genes (rows in X) representing a **connected** component in A (→ an altered sub-network) to predict the class labels for new samples



GenePEN - Approach

Idea: Cast the gene selection as an optimization problem, maximizing two quantities:

- the diagnostic prediction accuracy of the classifier
- connectedness of selected genes in the network

→ use a mathematical programming formulation (details on next slide):

$$\min_{w} \underline{loss(w)} + \underline{\lambda} \cdot \underline{penalty(w)}$$

loss-function (minimize error)   trade-off parameter   penalty-function (gene grouping)

→ Output: an optimized vector of feature weights ***w**:*

$\mathbf{w}_i \approx 0 \rightarrow$ gene i not selected

abs($\mathbf{w}_i$) large $\rightarrow$ gene is relevant for the prediction and well-grouped with other selected genes in the network

GenePEN – The loss and penalty functions

GenePEN objective function:

$$\min_{w} \underbrace{loss(w)}_{(1)} + \lambda \cdot \underbrace{penalty(w)}_{(2)}$$

- **(1)** the loss function is the expected logistic loss (smooth and convex → can be minimized efficiently):

$$loss(\underbrace{w}_{\text{gene weights}}, \underbrace{\nu}_{\text{offset parameter}}) = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \log\left(1+\exp(-\underbrace{y_i}_{\text{real labels}}\underbrace{(w^\top x_i+\nu)}_{\text{predicted labels}})\right)$$

- **(2)** the new convex penalty function penalizes the differences of absolute values (= measure of relevance) between the weights of neighboring genes/proteins:

$$penalty(\omega) = \sum_{i=1}^{p}\left[\sum_{j=1}^{p} \underbrace{A_{ij}}_{\text{adjacency matrix}} \left(|w_i| - |w_j|\right)\right]^2 + 2\underbrace{\Delta}_{\text{maximum network degree}} \|\omega\|_1^2$$

Previous penalty functions proposed

$\Omega(\mathbf{w}) = \|\mathbf{w}\|_2^2$	**Ridge** (Hoerl and Kennard, 1970) grouping but no sparsity
$\Omega(\mathbf{w}) = \|\mathbf{w}\|_1$	**Lasso** (Tibshirani, 1996) sparsity but no grouping
$\Omega(\mathbf{w}) = \|\mathbf{w}\|_2^2 + \alpha\|\mathbf{w}\|_1$	**Elastic Net** (Zou and Hastie, 2005) cannot capture local structure
$\Omega(\mathbf{w}) = \sum_{c\in\mathcal{C}} \alpha_c \|\mathbf{w}_c\|_2$	**Group Lasso** (Turlach et al., 2005) assumes non-overlapping groups
$\Omega(\mathbf{w}) = \mathbf{w}^\intercal\, \mathbf{K}\, \mathbf{w}$ (with $\mathbf{K}$ psd)	**graph kernel** (Rapaport et al., 2007) weight signs can introduce bias
$\Omega(\mathbf{w}) = \sum_{i<j} \max(\|w_i\|, \|w_j\|)$	**OSCAR** (Bondell and Reich, 2008) large weights can introduce bias

GenePEN – Application to Parkinson‘ disease data

- **Parkinson’s disease test dataset**: Microarray gene expression data from *post mortem* brain samples (*substantia nigra*) of 43 PD patients and 50 controls (Zhang et al., 2005)

- **Network data**: Human genome-scale protein-protein interaction network constructed from 80,543 public, direct physical interactions between 10,042 proteins.

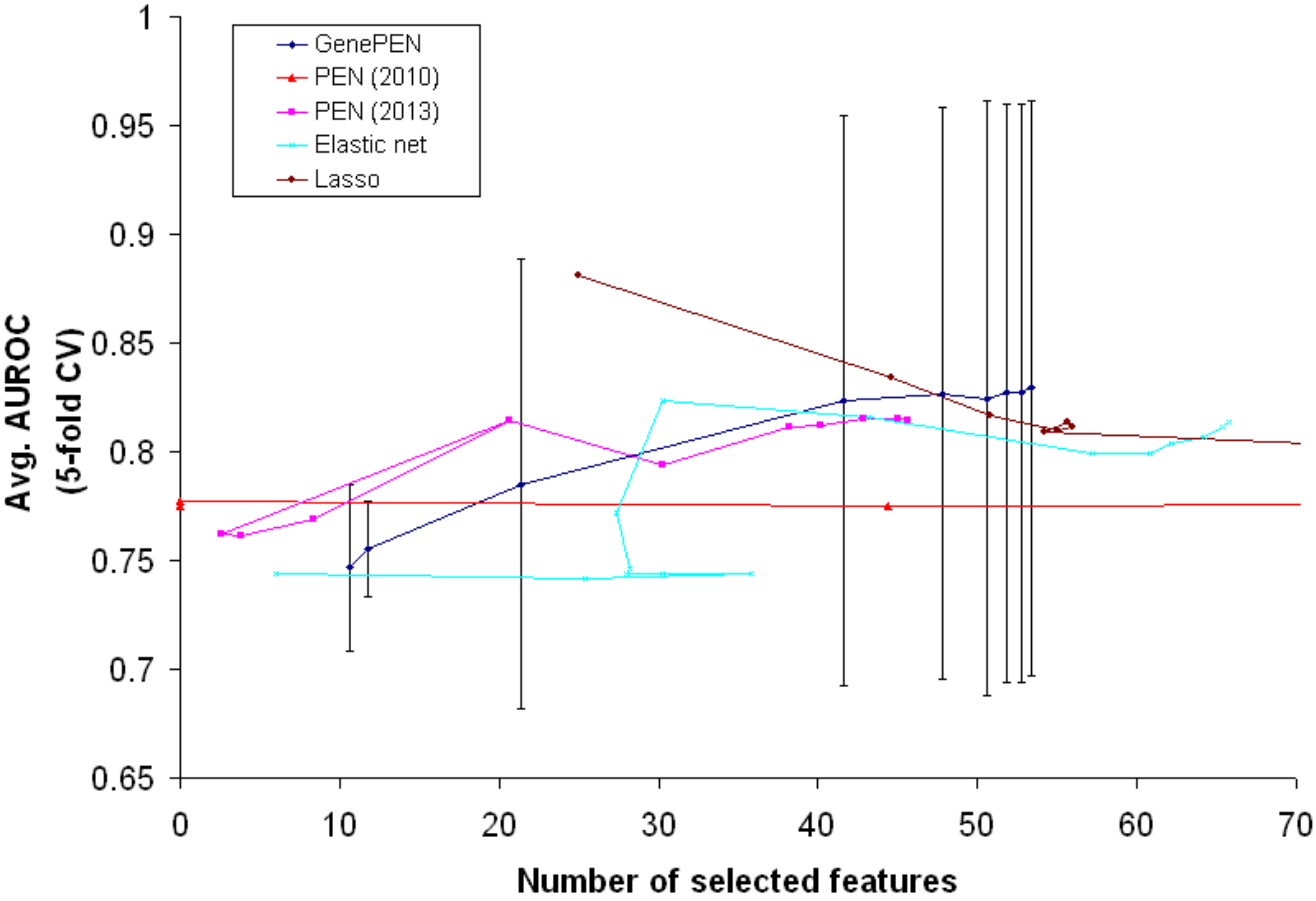
- **Comparison against other penalty functions**: The GenePEN penalty was compared against alternative penalty functions (Lasso, Elastic Net, Pairwise Elastic Net)

- **Evaluation criteria**:

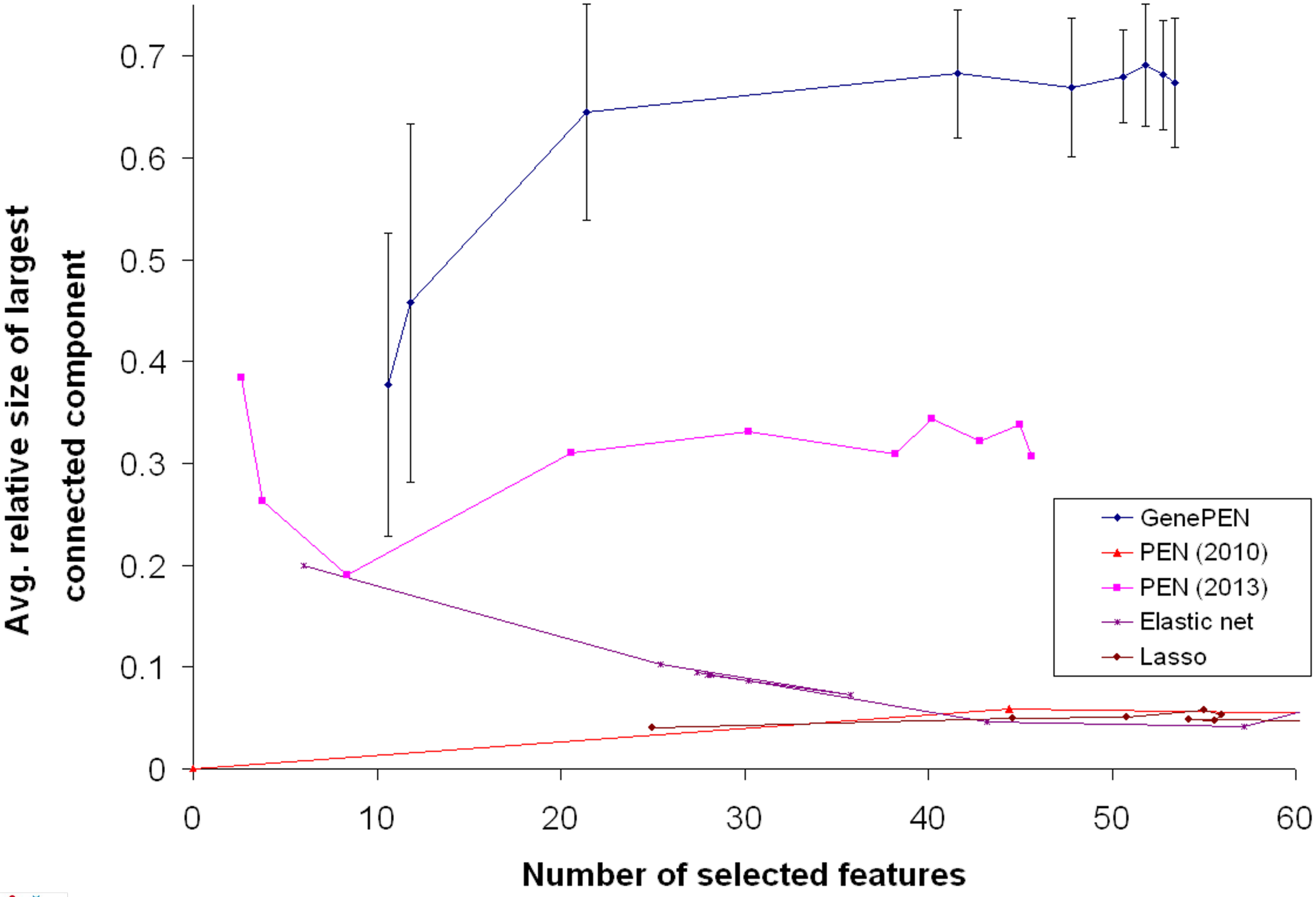
  → **cross-validated prediction performance**:
  avg. area under the receiver operating characteristic curve (AUROC) for different numbers of selected features

  → **cross-validated grouping of selected genes in the network**:
  avg. relative size of the largest connected component among selected features in the network

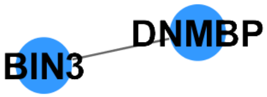
Comparison of AUROC performance



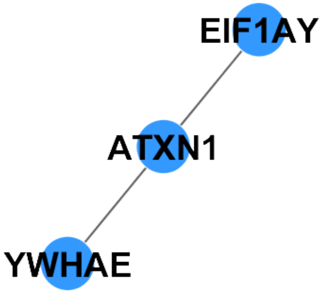
Relative size of largest connected component in network



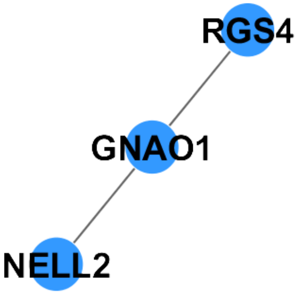
Comparison: Largest cluster for ~50 selected genes



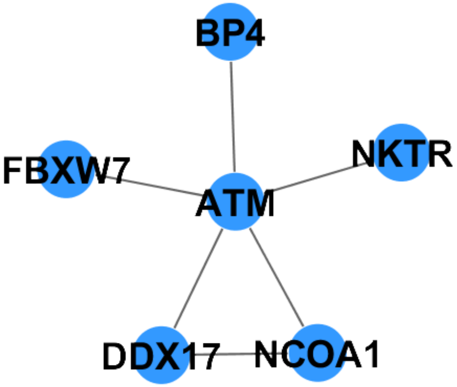
Lasso



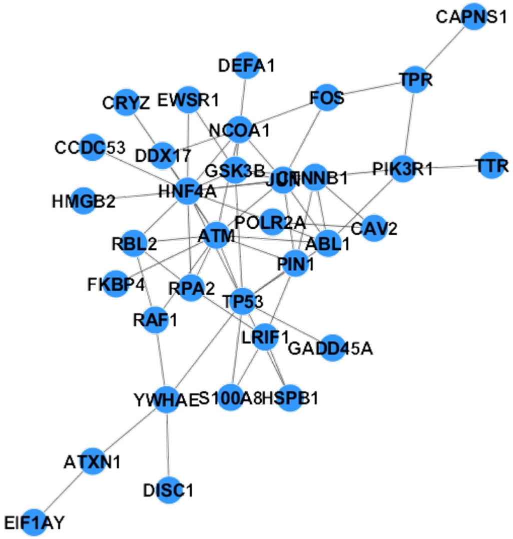
Elastic Net



PEN (2010)



PEN (2013)



GenePEN → cluster of 34 genes

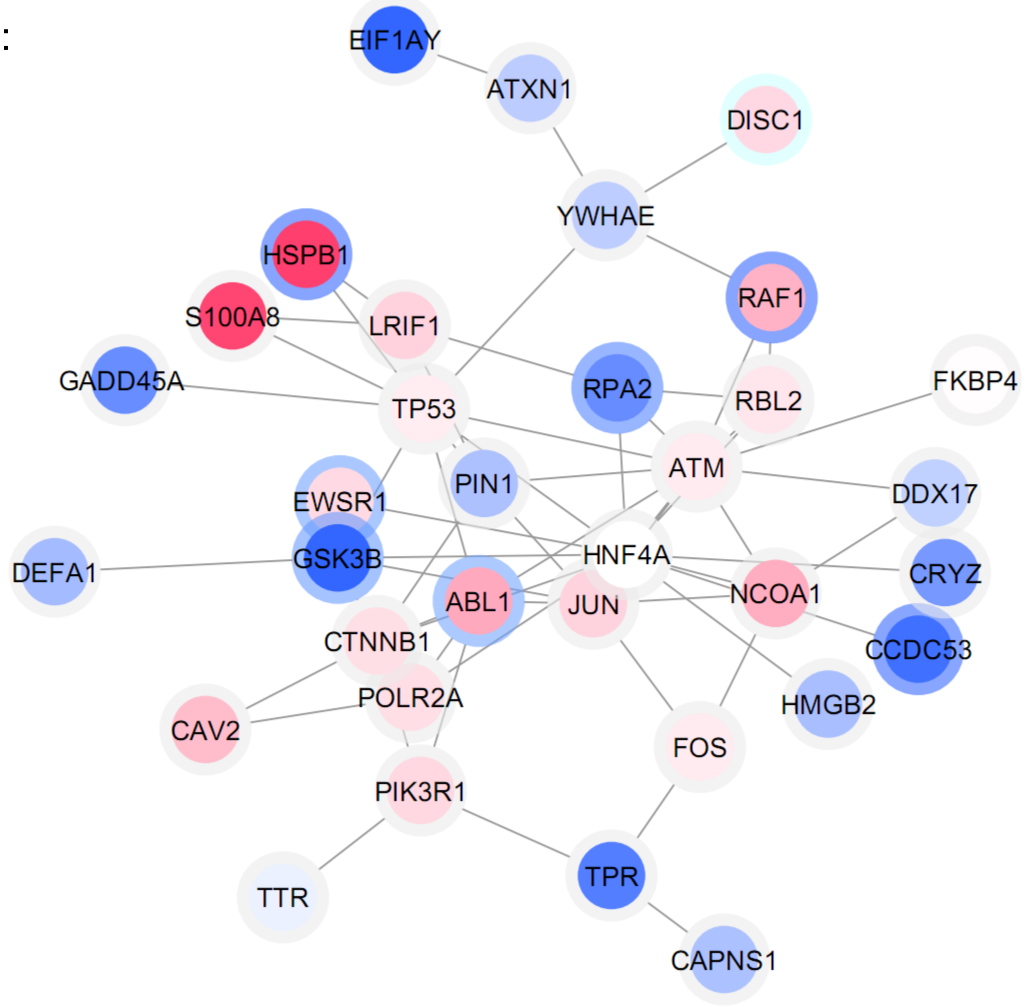
Biological results: PD-associated sub-network

Largest connected graph component identified on PD transcriptomics data:

- red = over-expressed in PD
  blue = under-expressed in PD
  node borders = individual statistical significance (from gray to blue with increasing significance)

- individually significant genes are significantly over-represented in the sub-network (p = 0.01)

- Pointwise mutual information (PMI) co-occurrence scoring of gene names and MeSH disease term "Parkinson's disease" in PubMed reveals enrichment of positive scores



Summary & Acknowledgements

- Integrating **prior knowledge** from molecular networks and pathways into omics data analysis can provide benefits in terms of model robustness and biological interpretability

- **GenePEN** discovers **discriminative sub-networks** for diagnostic sample classification and enables an interpretation of disease-associated molecular alterations at the network level

- On **Parkinson‘s disease transcriptomics data** GenePEN identifies predictive alterations in sub-networks which are enriched in individually significant genes and known PD-associated genes with positive PMI scores

Acknowledgments
Alex Lorbert, Martin Jaggi, Rudi Balling

Luxembourg
National
Research Fund

Publications

1. N. Vlassis, E. Glaab, *GenePEN: analysis of network activity alterations in complex diseases via the pairwise elastic net*, Statistical Applications in Genetics and Molecular Biology (2015), 14(2), 221
2. A. Rauschenberger, Z. Landoulsi, M. A. van de Wiel, E. Glaab. *Penalized regression with multiple sources of prior effects*, Bioinformatics (2022), 39(12), doi: 10.1007/s12035-022-02985-2.
3. M. Ali, O. Uriarte Huarte, T. Heurtaux, P. Garcia, B. Pardo Rodriguez, K. Grzyb, R. Halder, A. Skupin, M. Buttini, E. Glaab. *Single-Cell Transcriptional Profiling and Gene Regulatory Network Modeling in Tg2576 Mice Reveal Gender-Dependent Molecular Features Preceding Alzheimer-Like Pathologies*, Mol Neurobiol (2022), doi:10.1007/s12035-022-02985-2.
4. A. Rauschenberger, E. Glaab. *Predicting Dichotomised Outcomes from High-Dimensional Data in Biomedicine*, Journal of Applied Statistics, (2023), doi: 10.1080/02664763.2023.2233057.
5. L. C. Tranchevent, R. Halder, E. Glaab. *Systems level analysis of sex-dependent gene expression changes in Parkinson's disease*, NPJ Parkinson's Disease, (2022), 9, 8.
6. A. Rauschenberger, E. Glaab, *Predicting correlated outcomes from molecular data*, Bioinformatics (2021), 37(21), 3889–3895
7. R. Diaz-Uriarte, E. Gómez de Lope, R. Giugno, H. Fröhlich, P. V. Nazarov, I. A. Nepomuceno-Chamorro, A. Rauschenberger, E. Glaab, *Ten Quick Tips for Biomarker Discovery and Validation Analyses Using Machine Learning*, PLoS Computational Biology (2022), doi:10.1371/journal.pcbi.1010357
8. E. Glaab, J.P. Trezzi, A. Greuel, C. Jäger, Z. Hodak, A. Drzezga, L. Timmermann, M. Tittgemeyer, N. J. Diederich, C. Eggers, Integrative analysis of blood metabolomics and PET brain neuroimaging data for Parkinson's disease, Neurobiology of Disease (2019), Vol. 124, No. 1, pp. 555
9. S. Köglsberger, M. L. Cordero-Maldonado, P. Antony, J. I. Forster, P. Garcia, M. Buttini, A. Crawford, E. Glaab, *Gender-specific expression of ubiquitin-specific peptidase 9 modulates tau expression and phosphorylation: possible implications for tauopathies*, Molecular Neurobiology (2017), 54(10), pp. 7979
10. E. Glaab, *Using prior knowledge from cellular pathways and molecular networks for diagnostic specimen classification*, Briefings in Bioinformatics (2015), 17(3), pp. 440
11. E. Glaab, R. Schneider, *Comparative pathway and network analysis of brain transcriptome changes during adult aging and in Parkinson's disease*, Neurobiology of Disease (2015), 74, 1-13
12. E. Glaab, R. Schneider, *RepExplore: Addressing technical replicate variance in proteomics and metabolomics data analysis*, Bioinformatics (2015), 31(13), pp. 2235
13. E. Glaab, A. Baudot, N. Krasnogor, A. Valencia. Extending pathways and processes using molecular interaction networks to analyse cancer genome data, BMC Bioinformatics, 11(1):597, 2010
14. E. Glaab, A. Baudot, N. Krasnogor, R. Schneider, A. Valencia. EnrichNet: network-based gene set enrichment analysis, Bioinformatics, 28(18):i451-i457, 2012
15. E. Glaab, G. B. Manoharan, D. Abankwa, A Pharmacophore Model for SARS-CoV-2 3CLpro Small Molecule Inhibitors and in Vitro Experimental Validation of Computationally Screened Inhibitors, Journal of Chemical Information and Modeling (2021), 61(8), 4082–4096
16. E. Glaab, A. Rauschenberger, R. Banzi, C. Gerardi, P. Garcia, J. Demotes-Mainard, and the PERMIT Group, Biomarker discovery studies for patient stratification using machine learning analysis of omics data: a scoping review, BMC Open (2021), 11, e053674
17. D. M. Hendrickx, P. Garcia, A. Ashrafi, A. Sciortino, K. J. Schmit, H. Kollmus, N. Nicot, T. Kaoma, L. Vallar, M. Buttini, E. Glaab, A new synuclein-transgenic mouse model for early Parkinson's reveals molecular features of preclinical disease, Molecular Neurobiology (2020), 58, 576-602