PERMIT Workshop

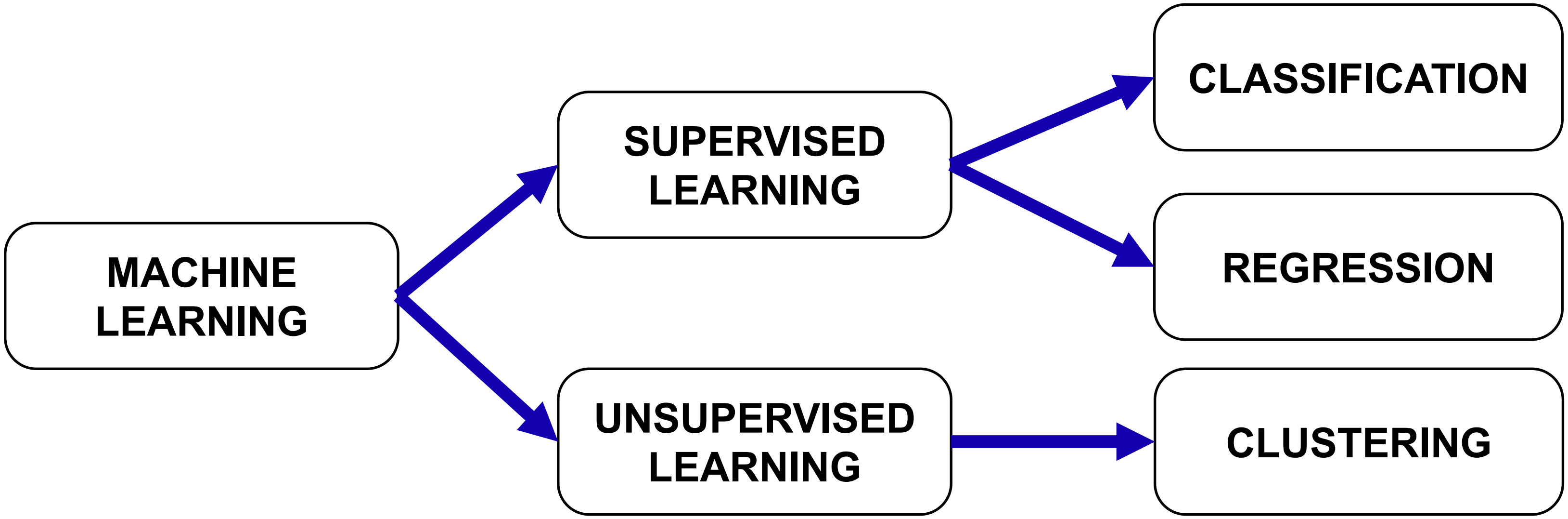
Scoping review on machine learning methods for stratification

Enrico Glaab
Luxembourg Centre for Systems Biomedicine

Scoping Review Objectives

GOALS:

- Inventory AI methods for omics-based patient stratification and their validation (supervised and unsupervised ML approaches)
- Identify limitations, challenges, gaps and existing recommendations



Research Questions

Machine learning methods for stratification:

- What are the main types of supervised and unsupervised ML methods for omics-based stratification? What are the recommended workflows?
- What are the specific strengths/weaknesses of different stratification approaches?

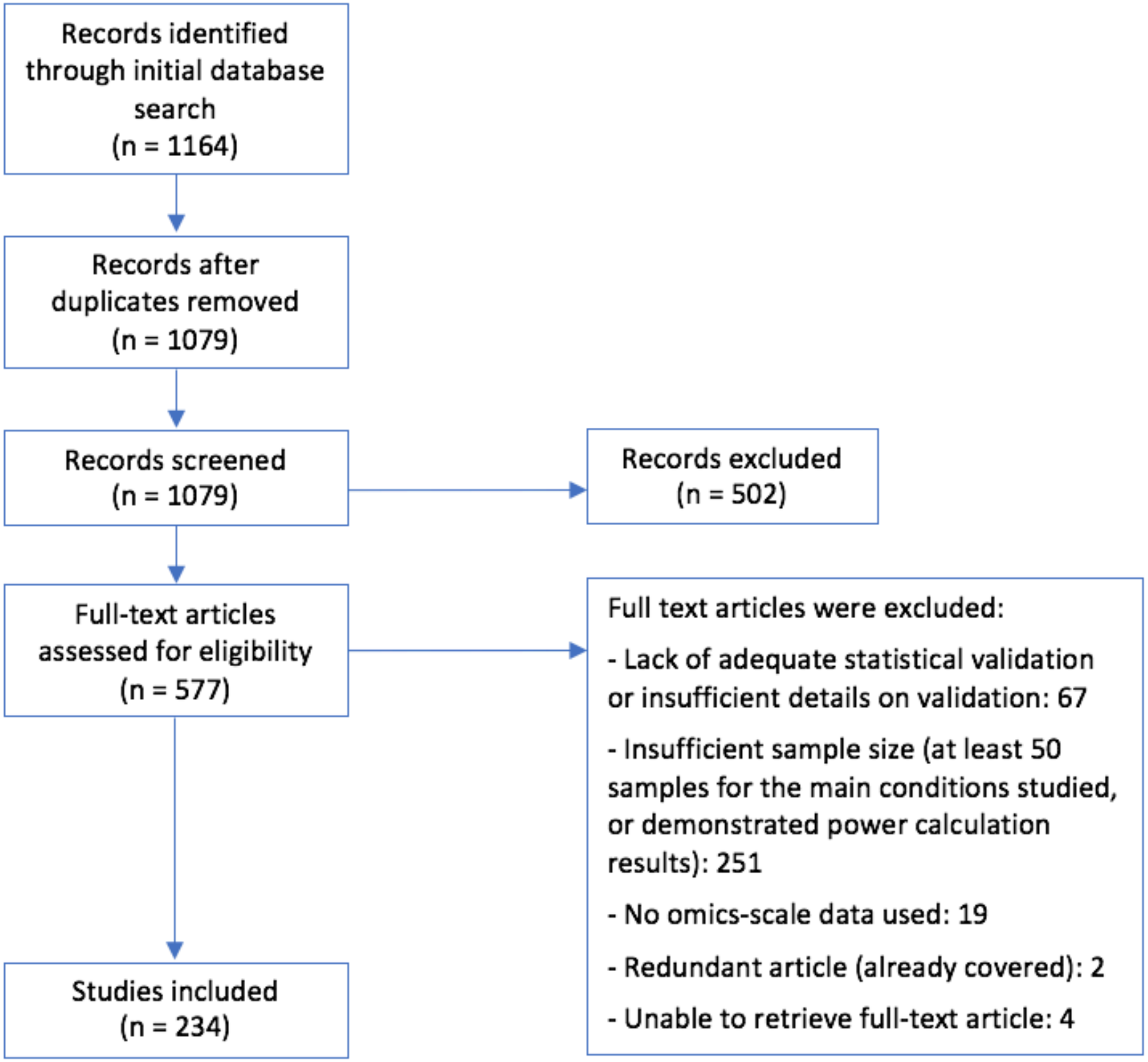
Validation methods:

- Which validation methods are available to assess accuracy, robustness and biomedical relevance? What are their strengths/weaknesses?

Applications:

- Which practical utility has been demonstrated in real-world settings (success/failure stories, lessons learned)?

Scoping Review Results: PRISMA flow diagram



Records identified through initial database search (n = 1164)

Records after duplicates removed (n = 1079)

Records screened (n = 1079)

Records excluded (n = 502)

Full-text articles assessed for eligibility (n = 577)

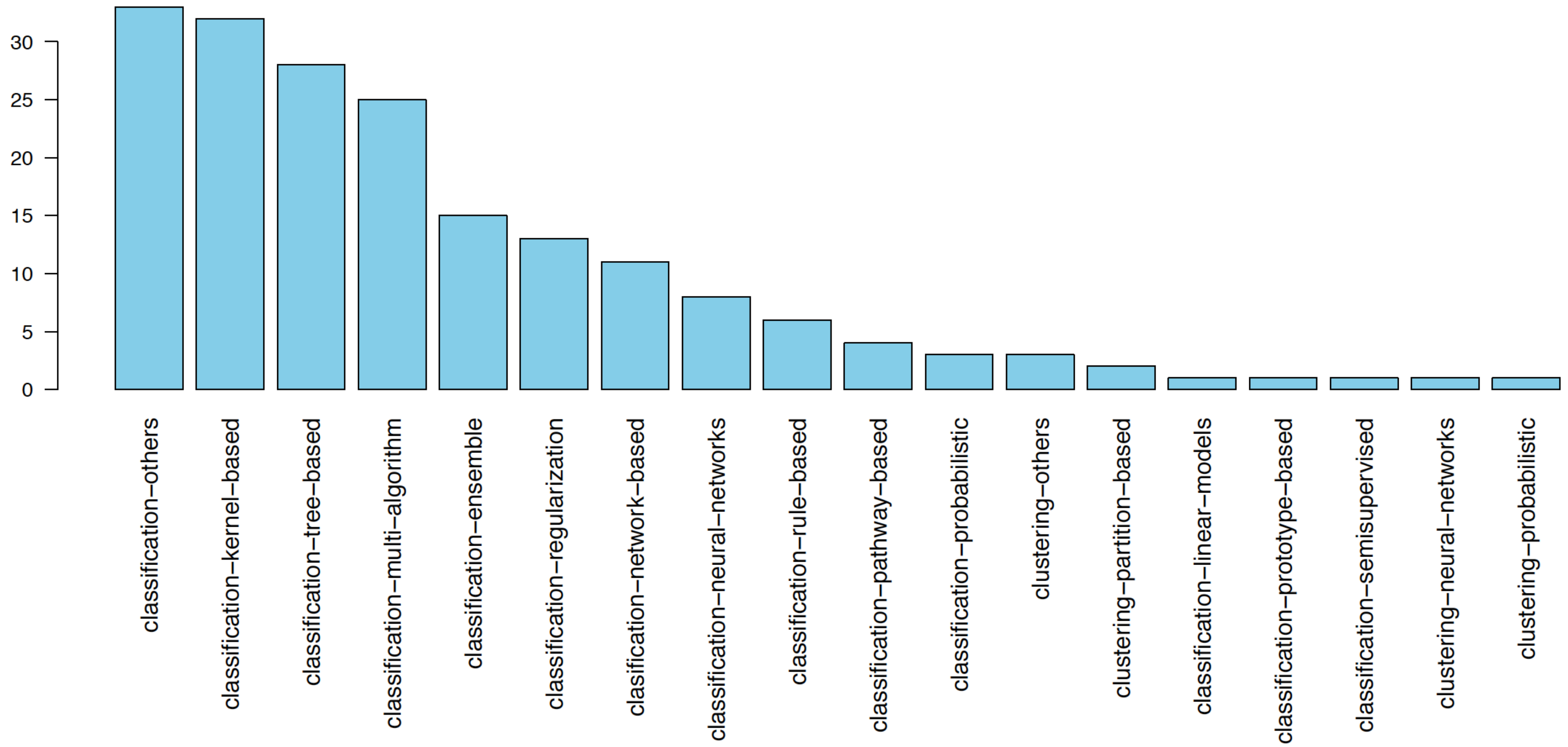
Full text articles were excluded:

- Lack of adequate statistical validation or insufficient details on validation: 67
- Insufficient sample size (at least 50 samples for the main conditions studied, or demonstrated power calculation results): 251
- No omics-scale data used: 19
- Redundant article (already covered): 2
- Unable to retrieve full-text article: 4

Studies included (n = 234)

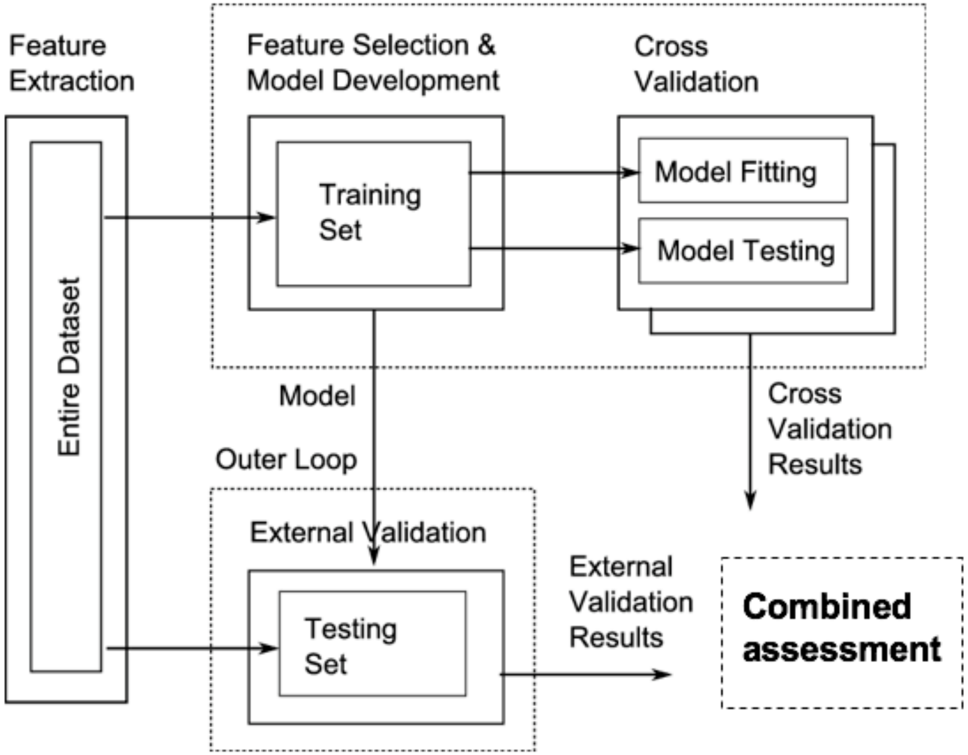
Results: Used ML methodologies

- Over-represented approaches: Tree- and kernel-based ML methods
- Under-represented approaches: Probabilistic, prototype-based and neural network based ML methods



Results: Used validation methodologies

- Common methods: Training/test set split + cross-validation on training data (LOOCV, 10-fold CV), metrics: accuracy, AUC
- Less frequent: External cohort validation, robust bootstrapping and & bolstered CV approaches, metrics: F1, MCC, PR-AUC



Main gaps & limitations identified (1)

Study design and documentation related issues:

(1) study group design and sample size selection

(underpowered, imbalanced)

(2) statistical evaluation

(robustness/completeness, multiple hypothesis testing)

(3) clarity of clinical applications

(primary/secondary outcomes)

(4) study documentation

(settings/parameters, reproducibility)

Main gaps & limitations identified (2)

Issues affecting model reliability, robustness and interpretability:

(5) Sampling & blocking design
(batch effects and biases)

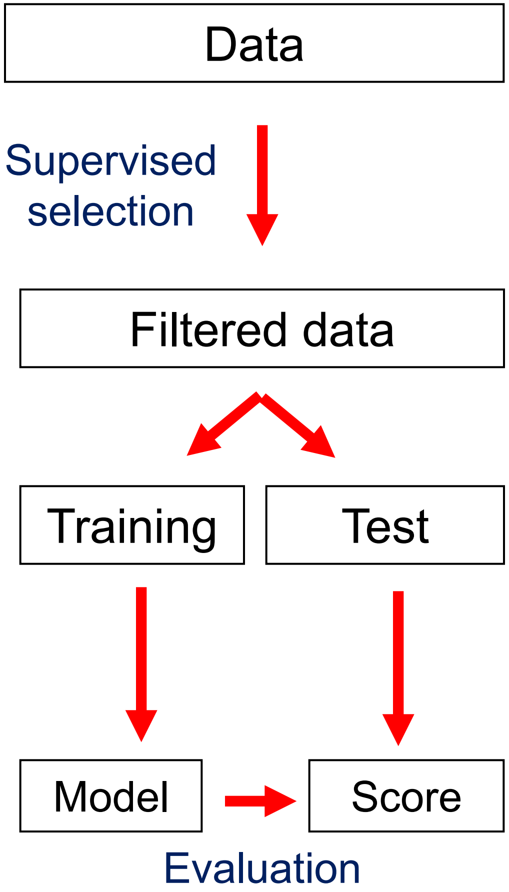
(6) data pre-processing, filtering and normalization
(lacking standards)

(7) integration of prior biological knowledge
(pathway/network knowledge, multi-omics analyses)

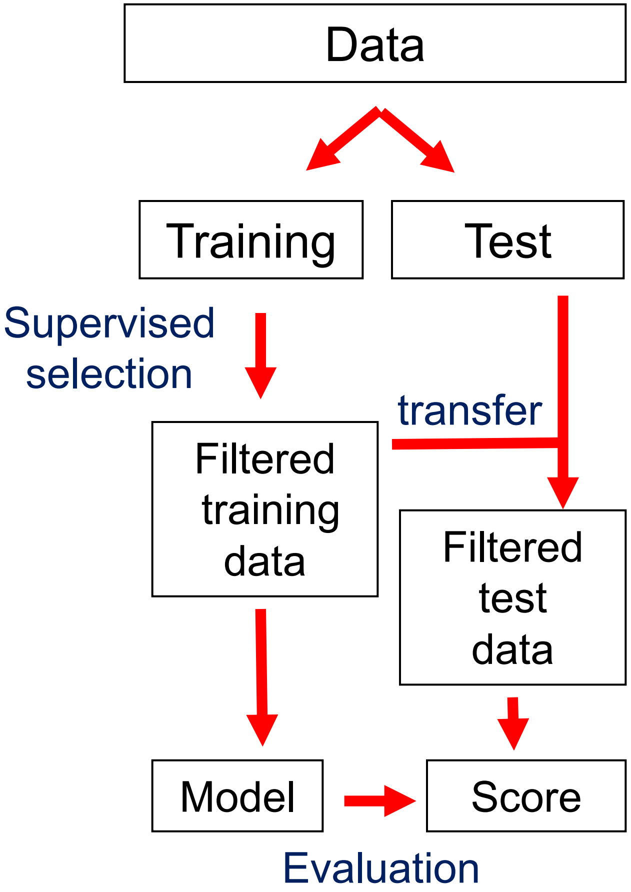
(8) Ensuring model interpretability and biological plausibility
(black-box vs. white-box models)

Gaps & limitations: Example

Common error: Global feature selection



Correct approach



Recommendations from the scoping review / literature

Data pre-processing, filtering & normalization:

→ use cross-validation to check if pre-processing leads to information loss

→ compare or combine multiple pre-processing approaches

Integration of prior knowledge & multi-omics analyses:

→ check prior literature on the cost/benefit of multi-omics analyses for the studied conditions / cell types, or conduct pilot analyses

→ use existing software & frameworks for integrative biological data analysis

Ensuring model interpretability & biological plausibility:

→ use dedicated methods to build interpretable models (e.g. rule learning)

→ use cellular pathway/network analysis & literature mining to guide modeling

Previous success stories

- Multiple omics-derived biomarker signatures already clinically validated
- Most tests are for cancer diseases, but first non-cancer applications exist

Name	Test approval (FDA-cleared and/or LDT)	Purpose	References
MammaPrint	FDA-cleared, LDT	breast cancer risk-of-recurrence assessment	Van't Veer et al., Nature, 2002
AlloMap Heart	FDA-cleared, LDT	identifying heart transplant recipients with risk of cellular rejection	Yamani et al., J Heart Lung Transplant, 2007
Prosigna Assay / PAM50	FDA-cleared, LDT	breast cancer risk of distant recurrence prediction	Nielsen et al., BMC Cancer, 2014
Oncotype DX	LDT	breast cancer risk-of-recurrence assessment	Kelley et al., Cancer, 2010
Decipher	LDT	prostate cancer metastatic risk prediction	Marrone et al., PLoS Curr., 2015

Previous success stories – Main conclusions

Shared characteristics of prior success stories as a guideline:

- Early filtering:
  rigorous statistical, clinical and biological filtering criteria applied (strict inclusion/exclusion criteria; multiple layers of statistical and ML-based feature selection; integration of prior knowledge)

- Continuous technological improvements:
  transition from cheap, low-sensitivity to high-sensitivity measurements (e.g. from microarray technology to deep sequencing, RT-PCR and digital PCR)

- Robust validation schemes:
  multi-level cross-validation, bootstrapping and external validation involving multiple performance metrics, large sample sizes, and multiple cohorts

Summary

Main gaps & limitations:

→ study design: many studies are underpowered, imbalanced

→ statistical validation: often incomplete, lacking robustness or even incorrect

→ study documentation: lack of details, irreproducible

Main proposed recommendations:

→ follow existing study design & documentation guidelines (e.g. NCI check list)

→ use robust validation schemes & early filtering

→ exploit prior biological knowledge & existing data integration frameworks

References

1. E. Glaab, *Using prior knowledge from cellular pathways and molecular networks for diagnostic specimen classification*, Briefings in Bioinformatics (2015), 17(3), 440
2. 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
3. E. Glaab, J. M. Garibaldi, N. Krasnogor. *Learning pathway-based decision rules to classify microarray cancer samples*, German Conference on Bioinformatics 2010, Lecture Notes in Informatics (LNI), 173, 123-134
4. E. Glaab, J. Bacardit, J. M. Garibaldi, N. Krasnogor, *Using rule-based machine learning for candidate disease gene prioritization and sample classification of cancer gene expression data*, PLoS ONE, 7(7):e39932, 2012
5. 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
6. E. Glaab, A. Baudot, N. Krasnogor, R. Schneider, A. Valencia. EnrichNet: network-based gene set enrichment analysis, Bioinformatics, 28(18):i451-i457, 2012
7. Maes, M., Nowak, G., Caso, J. R., Leza, J. C., Song, C., Kubera, M., .et al. (2016). Toward omics-based, systems biomedicine, and path and drug discovery methodologies for depression-inflammation research. Molecular neurobiology, 53(5), 2927-2935.
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. 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
10. Z. Zhang, P. P. Jung, V. Grouès, P. May, C. Linster, E. Glaab, *Web-based QTL linkage analysis and bulk segregant analysis of yeast sequencing data*, GigaScience (2019), 8(6), 1-18
11. 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
12. Kleiderman, S., Gutbier, S., Ugur Tufekci, K., Ortega, F., Sá, J. V., Teixeira, A. P., et al. (2016). Conversion of Nonproliferating Astrocytes into Neurogenic Neural Stem Cells: Control by FGF2 and Interferon-γ. Stem Cells, 34(12), 2861-2874.
13. Bolognin, S., Fossépré, M., Qing, X., Jarazo, J., Ščančar, J., Moreno, E. L., et al. (2019). 3D Cultures of Parkinson's Disease-Specific Dopaminergic Neurons for High Content Phenotyping and Drug Testing. Advanced Science, 6(1), 1800927.
14. Jaeger, C., Glaab, E., Michelucci, A., Binz, T. M., Koeglsberger, S., Garcia, P., ... & Buttini, M. (2015). The mouse brain metabolome: region-specific signatures and response to excitotoxic neuronal injury. The American Journal of Pathology, 185(6), 1699-1712.
15. E. Glaab, R. Schneider, *RepExplore: Addressing technical replicate variance in proteomics and metabolomics data analysis*, Bioinformatics (2015), 31(13), pp. 2235
16. E. Glaab, *Building a virtual ligand screening pipeline using free software: a survey*, Briefings in Bioinformatics (2015), 17(2), pp. 352
17. E. Glaab, R. Schneider, *PathVar: analysis of gene and protein expression variance in cellular pathways using microarray data*, Bioinformatics, 28(3):446-447, 2012
18. E. Glaab, A. Baudot, N. Krasnogor, A. Valencia. *TopoGSA: network topological gene set analysis*, Bioinformatics, 26(9):1271-1272, 2010
19. E. Glaab, J. M. Garibaldi and N. Krasnogor. *ArrayMining: a modular web-application for microarray analysis combining ensemble and consensus methods with cross-study normalization*, BMC Bioinformatics,10:358, 2009