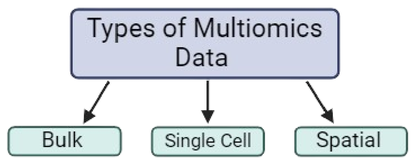
MESSI : Multiomics Experiments with SyStematic Interrogation

Chunqing Tony Liang [1,3] , Amrit Singh [2,3]

[1] UBC: Faculty of Science, Bioinformatics , [2] Faculty of Medicine, Anesthesiology, Pharmacology and Therapeutics, [3]Centre for Heart Lung Innovation

INTRODUCTION

- **Omics** is the comprehensive study of all molecules of a particular type within an organism (**e.g. proteins, metabolites, genes**)
- Multiomics compensates for missing or unreliable information in any single omics data



DATA INTEGRATION

Combine individual omics data, in a sequential or simultaneous manner, to understand the interplay of molecules.

- Same N, different P (N-integration)
- Sample P, different N (P-integration)

OBJECTIVES

The objective is to benchmark methods for multiomics data integration by:

1. Curating publicly available multiomics data
2. Applying existing methods to simulated and real world datasets
3. Compare methods based on: **classification accuracy, features selected & computation time**

MOTIVATIONS

- Many methods: **which to use?**
- How to **reproduce** method and get same results?
- Create a benchmarking study that is **all encompassing** (many methods and datasets)

METHODS

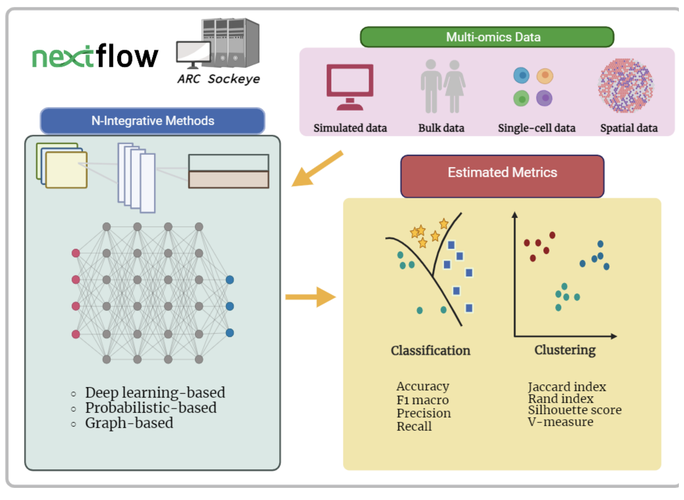


Figure 1: Overview of MESSI Workflow. Data are standardized into MuData and MultiAssayExperiemnt, then passed to methods for cross-validation in a parallel and fully isolated through processes and containers. Then each performance will be assessed by its task (classification or clustering)corresponding metrics.

RESULTS

Datasets	Outcome	Number of patients	Number of Omics	Disease	Source
mixOmics Breast	Basal/non-basal	150	3	Breast Cancer	The Cancer Genome Atlas
ROSMAP	AD/non-AD	351	3	Alzheimers	NIAGADS Data Sharing Service
GSE71669	Invasive/non-invasive	33	3	Bladder Cancer	GEO
Simulated 1	cancerous/non-cancerous	100	5	Cancer A	-
Simulated 2	cancerous/non-cancerous	50	4	Cancer B	-
Simulated 3	cancerous/non-cancerous	30	3	Cancer C	-

Table 1: Dataset Characteristics. Summary of metadata from benchmark studies of interests. 3 datasets are real data, 3 are simulated. Task of these data are binary classification on cancer.

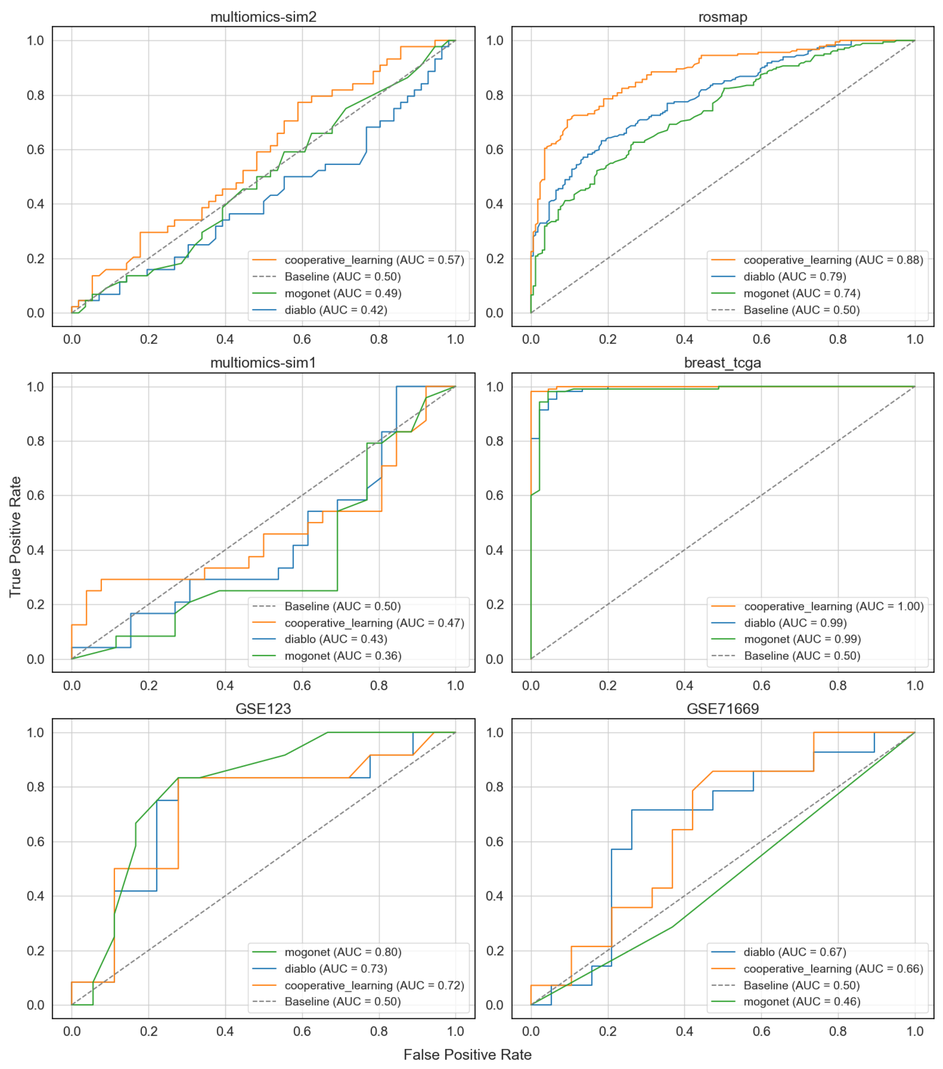


Figure 2: Receiver Operating Curves for each dataset VS method. ROC curves presented to show classification performance of each method in various dataset (real and simulated)

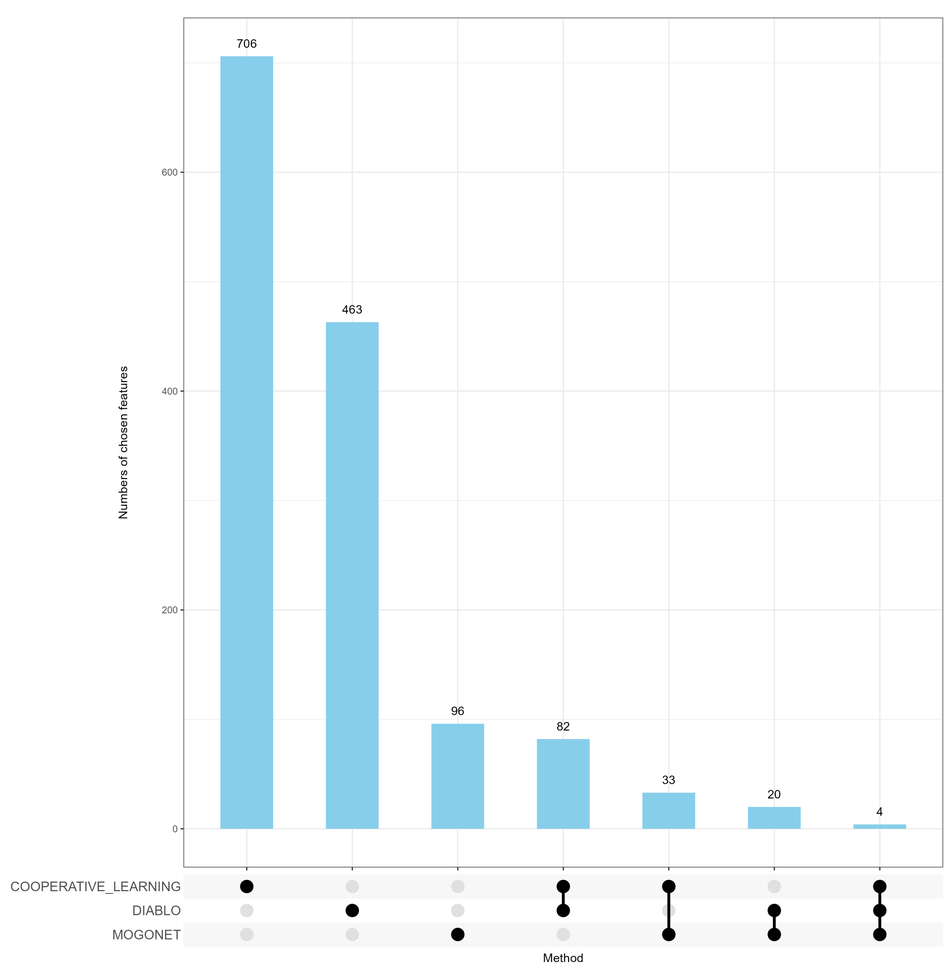


Figure 3: Upset plot of overlapped features selected per method. Bars denote size of features counts, and dots with line meaning if consistent in methods.

RESULTS

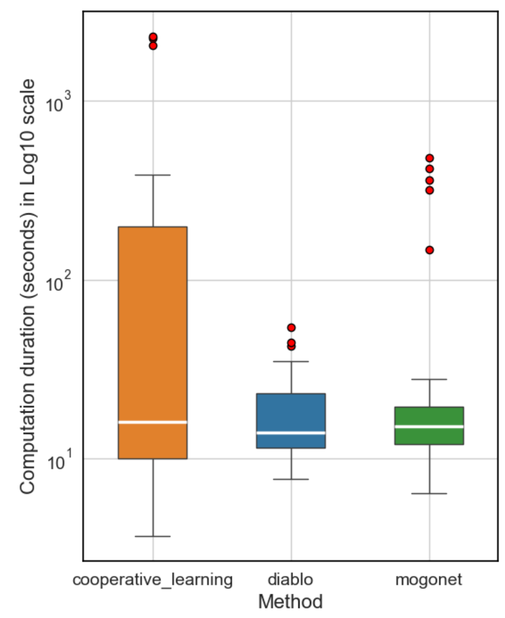


Figure 4: Duration of methods. Computation time of that a method takes from preprocess, train and predict all datasets.

CONCLUSION

- **No method works** universally well at simulated data with noise, often worse than random guessing
- **GNN like MOGONET, overcomplicated** yet low performance
- **DIABLO is the best** at moment, faster run time
- Need to generalize for more datasets, at sc or spatial level
- Add in other methods (like bayesian, other DLs)

REFERENCES

- Di Tommaso, Paolo, et al. "Nextflow enables reproducible computational workflows." *Nature biotechnology* 35.4 (2017): 316-319.
- Ding, Daisy Yi, et al. "Cooperative learning for multiview analysis." *Proceedings of the National Academy of Sciences* 119.38 (2022): e2202113119.
- Wang, Tongxin, et al. "MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification." *Nature communications* 12.1 (2021): 3445.
- Singh, Amrit, et al. "DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays." *Bioinformatics* 35.17 (2019): 3055-3062.

ACKNOWLEDGEMENTS

I would like to thank all members of the CompBio Lab at UBC, the Centre for Heart Lung Innovation, UBC Bioinformatics, and Providence healthcare , Vancouver.