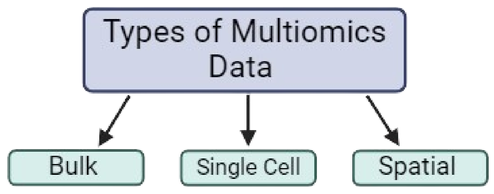
MESSI : Multiomics Experiments with SyStematic Interrogation

A Nextflow pipeline for benchmarking multiomics integration methods

Chunqing Tony Liang [1,3] , Tajveer Grewal[2], Asees Singh[2], Amrit Singh [1,2,3]
[1] 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



- Many integration methods: which to use?
- How to reproduce method and get same results?
- We introduce *Multiomics Experiments with SyStematic Interrogation* (MESSI).
- Tool to assist in the development and evaluation of new/existing multiomics integration methods.

METHODS

- Looking into multiview (3), DIABLO (4), RGCCA (5), MOFA (6), MOGONET (7)
- Using a **5-fold** cross validation (cv) with are under the curve (AUC) score as metric
- Validating with known ground truth data from **500+** simulated data with varying parameters
- Evaluate on 6 real world datasets

Dataset	# of subjects	Disease	Prop(Y=1)	Omics	# of predictors
GSE71669	33	Bladder cancer	0.424	mrna	5831
				cpg	8915
				cc	10
ROSMAP	351	Alzheimer Disease	0.519	epigenomics	58
				genomics	74
				transcriptomics	90
TCGA-BLCA	336	Bladder cancer	0.685	Methylation_HM450K	8072
				miRNA	285
				RPPA	44
				RNAseq_HiSeq	6376
TCGA-BRCA	109	Breast cancer	0.358	Methylation_HM450K	7985
				miRNA_GA	282
				RPPA	43
				RNAseq_HiSeq	6191
TCGA-KIPAN	123	Kidney cancer	0.520	Methylation_HM450K	7914
				miRNA_GA_miR	225
				RPPA	60
				RNAseq_HiSeq	6145
TCGA-THCA	217	Thyroid cancer	0.318	Methylation_HM450K	6367
				miRNA	256
				RPPA	25
				RNAseq_HiSeq	5692

Table1: Real world dataset description. Each dataset is a public source multiomics dataset that studies one disease. Prop(Y=1) represents the proportion of positive cases in the dataset. # of predictors are number of molecular measurements in individual omics. # of subjects represent the size of study performed.

RESULTS

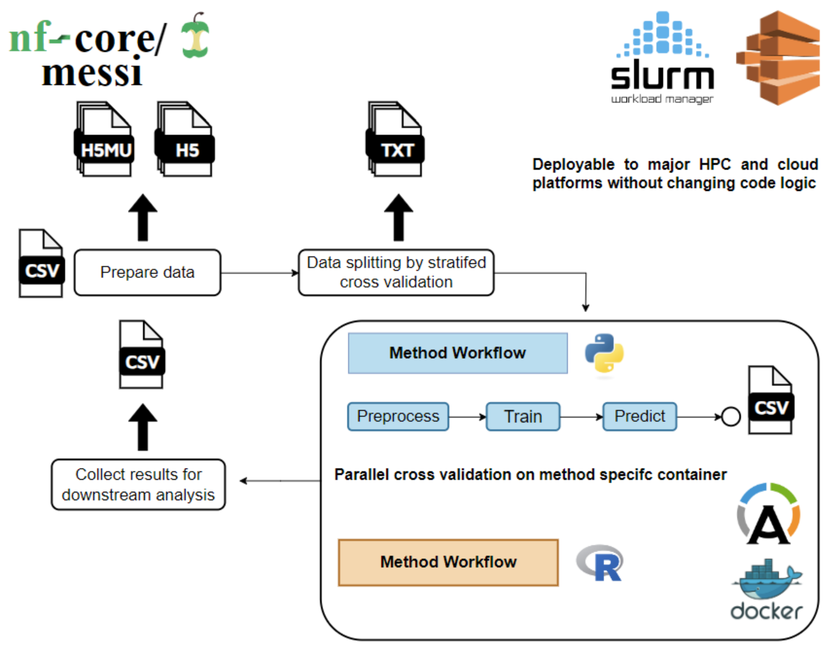


Figure 1: Overview of MESSI Workflow. Data are standardized into MuData and MultiAssayExperiemnt, then passed to methods for cross-validation in a parallel and fully isolated through processes and containers. Then each performance will be assessed by its task (classification or clustering) corresponding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

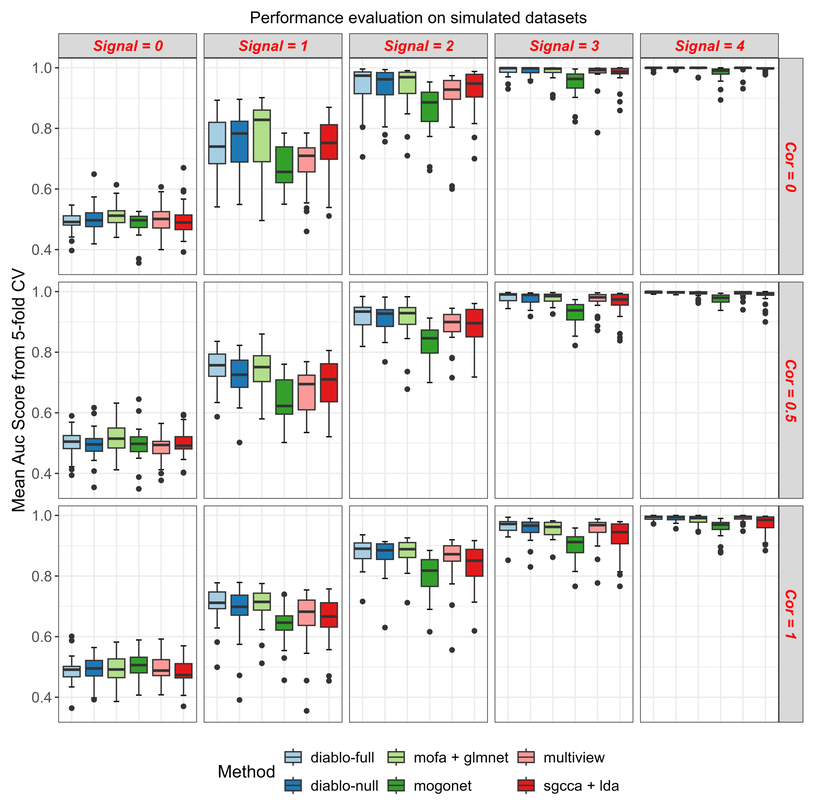


Figure 2: Evaluation of simulated datasets with varying parameters. Boxplot represents mean AUC score distributions from 5-fold CV, with controlled signal and correlation inside the datasets.

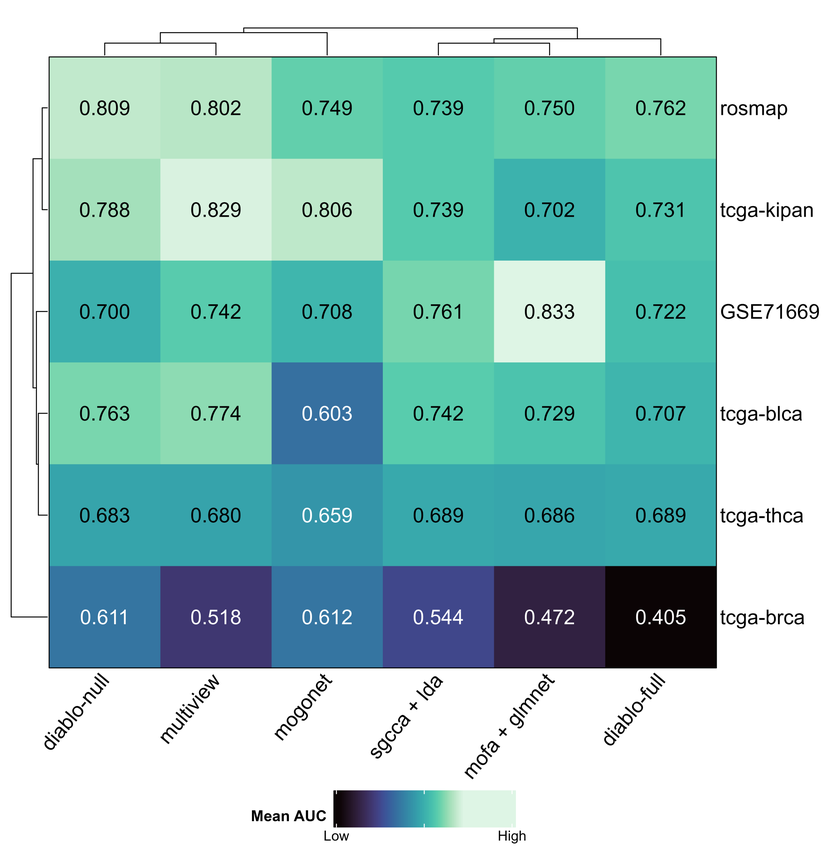


Figure 3: Performance of methods over real world datasets. Lighter color indicates better AUC score, numbers are exact score.

RESULTS

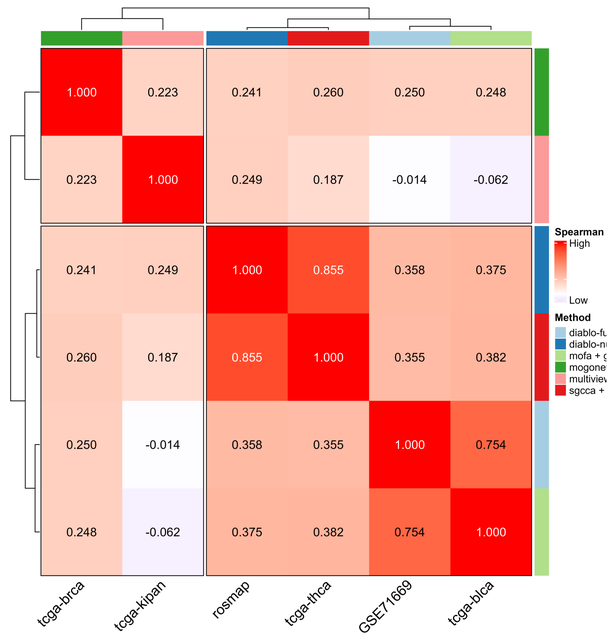


Figure 4: Spearman rank correlation on ranking biomarkers identified between methods. Darker color indicates higher similarity between the rankings of important biomarkers across method and dataset.

CONCLUSION

- **No method works** universally well on all datasets
- Pipeline proves way to reproducibly benchmark different aspects of integration methods
- GNN like **MOGONET,** overcomplicated yet low performance
- **DIABLO** is the best at moment
- Methods overall work better with higher effect, and less sensitive to correlation between omics
- Need to generalize for more datasets, at sc or spatial level
- Add in other methods (like bayesian, other DLs)

REFERENCES

1. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nature biotechnology. 2017;35(4):316–9.
2. Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, et al. The nf-core framework for community-curated bioinformatics pipelines. Nature biotechnology. 2020;38(3):276–8.
3. Ding DY, Li S, Narasimhan B, Tibshirani R. Cooperative learning for multiview analysis. Proceedings of the National Academy of Sciences. 2022;119(38):e2202113119.
4. Singh A, Shannon CP, Gautier B, Rohart F, Vacher M, Tebbutt SJ, et al. DIABLO: An integrative approach for identifying key molecular drivers from multi-omics assays. Bioinformatics. 2019;35(17):3055–62.
5. Girka F, Camenen E, Peltier C, Gloaguen A, Guillemot V, Le Brusquet L, et al. Multiblock data analysis with the RGCCA package. Journal of Statistical Software. 2023;1–36.
6. Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, et al. Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. Molecular systems biology. 2018;14(6):e8124.
7. Wang T, Shao W, Huang Z, Tang H, Zhang J, Ding Z, et al. MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. Nature Communications. 2021;12(1):3445.

ACKNOWLEDGEMENTS

THE UNIVERSITY OF BRITISH COLUMBIA
Department of Anesthesiology, Pharmacology & Therapeutics
Faculty of Medicine
Centre for Heart Lung Innovation
UBC and St. Paul's Hospital
CompBio Lab @ HLI
cbl-hli.med.ubc.ca/