Multi-omics data integration: Does more mean better for predictive modeling? A large-scale benchmark study

Yingxia Li[1], Tobias Herold[2], Ulrich Mansmann[1], **Roman Hornung**[1,3]

CEN Conference 2023, Basel

September, 7th, 2023

[1] **I**nstitute for Medical Information Processing, **B**iometry and **E**pidemiology, LMU Munich, Munich, Germany

[2] Department of Medicine III, University Hospital, LMU Munich, Munich, Germany

[3] Munich Center for Machine Learning (MCML), Munich, Germany

Introduction

- Multi-omics data are data for which **several types of omics data** (such as mRNA, miRNA, methylation data) are **available** for the **same patients**.

- In principle, **each** individual **type** of omics data **could be valuable** in the **prediction** of clinical outcomes.

$\Rightarrow$ The use of multi-omics data for prediction is becoming **increasingly popular**.

- The **prevailing view** in the literature seems to be that it is **best to include** as **many omics data types** as possible to maximize predictive performance.

Introduction

However, ...

- ... the predictive **information overlaps** between the different data types.
- ... the **level of** predictive **information varies** between data types.
- ... there are **interactions** between genomic features of **different data types**.

$\Rightarrow$ It is **possible that certain combinations** of data types perform generally **better** in prediction than others – **omitting** unnecessary data types **could reduce costs and efforts**.

$\Rightarrow$ We have conducted a **benchmark study** based on a large number of real data sets to determine **which combinations** of omics data types are generally **most effective** for prediction.

Study Design

- **14** multi-omics cancer **data sets** with **survival outcomes**; publicly available from The Cancer Genome Atlas (TCGA)
- **Data types** included: clinical features (`clin`), CNV (`cnv`), methylation (`met`), miRNA (`mirna`), mRNA (`rna`), mutation (`mut`)
- **Prediction modeling**:
  - **Methods**: LASSO (`lasso`), random survival forests (`rsf`), block forests (`bf`), IPF-LASSO (`ipflasso`), priority-LASSO (`prioritylasso`)
  - **Preliminary feature selection** for data types with more than 2,500 features
  - **Upweighting** of **clinical** features (not possible for `ipflasso`)

Study Design

- Performance **measures**: integrated Brier score (`ibrier`), Harrell's concordance index (`cindex`)
- Resampling/evaluation scheme: 5-fold **cross-validation** repeated five times
- For each data set and prediction method, **all $2^5 - 1 = 31$ combinations** of omics data types were considered; clinical features always included.
- **Bootstrap analysis** to examine the **stability** of the results ($B = 5000$)
- The **R code** implementing the benchmark study is **available on GitHub** (https://github.com/yingxiali/multi-omics-data).

Results: **Ranks** of the data type combinations (`ibrier`)

Multi-omics for prediction: more = better?

Roman Hornung

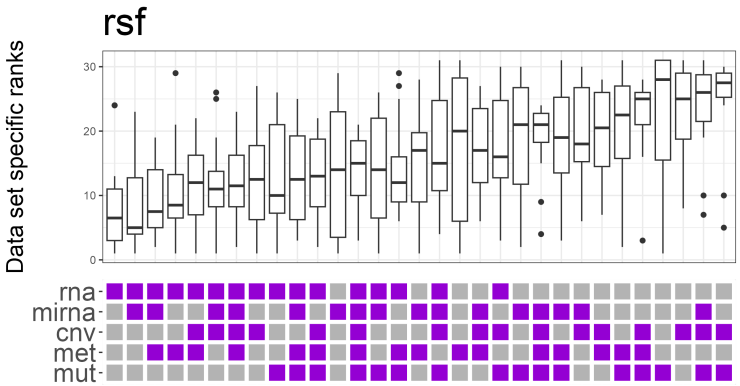
Introduction

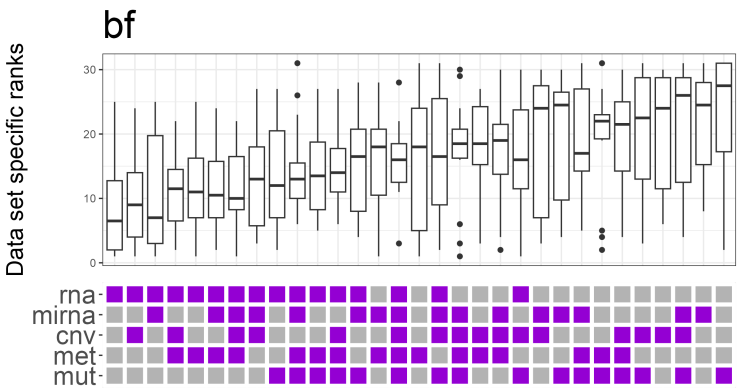
Study Design

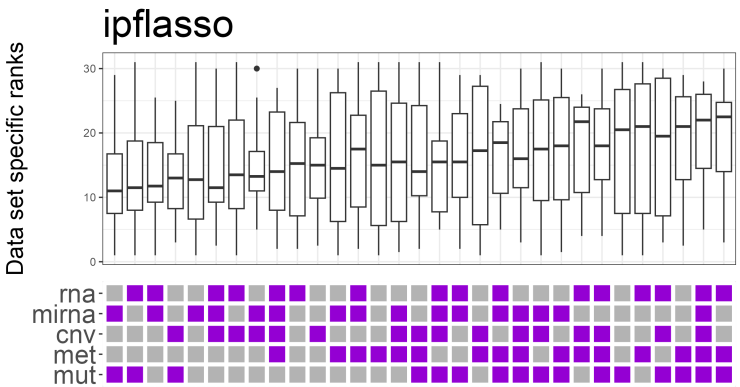
Results

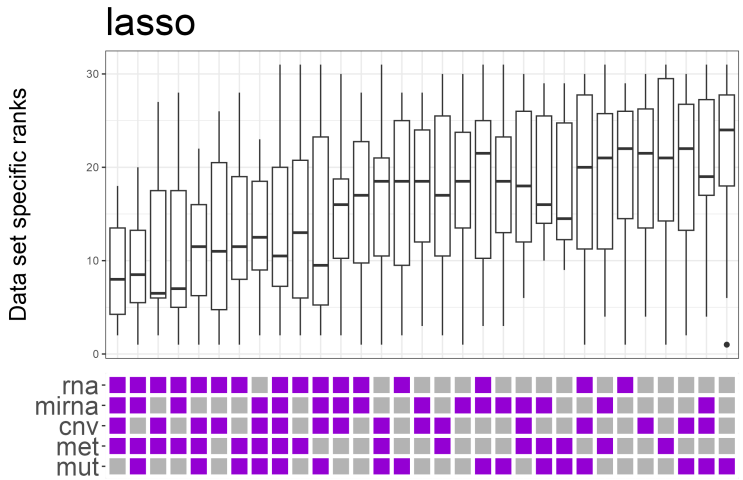
Discussion of Results

Conclusions









Results: **Best performing combinations** of prediction methods and data types per data set

	ibrier		cindex	
data set	prediction method	data type combination	prediction method	data type combination
BLCA	lasso	rna, mirna	prioritylasso	rna
BRCA	bf	rna, met	bf	rna, mirna
COAD	bf	met	bf	rna, mut
ESCA	ipflasso	mut	rsf	mirna, mut
HNSC	ipflasso	rna, mirna	rsf	rna
LGG	ipflasso	met, cnv	prioritylasso	rna
LIHC	bf	rna, mirna, met, cnv	lasso	rna, cnv
LUAD	lasso	mirna	ipflasso	mut
LUSC	ipflasso	mirna, met	prioritylasso	rna, mirna
PAAD	prioritylasso	rna	bf	rna
SARC	prioritylasso	met, cnv	rsf	mirna, met, mut
SKCM	lasso	rna, mut, cnv	bf	rna
STAD	rsf	rna, mirna, mut	rsf	mirna
UCEC	bf	rna, cnv	rsf	mirna

Discussion of Results

- **Strong variability** of **results** across data sets, prediction methods, and performance metrics
- **However**, **bootstrap analysis confirmed** that combining **few data types** for prediction tends to perform **better** than combining many.
- In particular, it confirmed that **using all available** data types does **not** yield **optimal** results.

Discussion of Results

- Large-scale benchmark study by **Herrmann et al. (2021)**: **discouraging results** regarding the **utility of multi-omics** data for prediction

- **Possible explanation**: Herrmann et al. **used all** available **data types**.

Conclusions

- While it is useful to analyze as many data types as possible in combination to understand cancer biology, only a **few data types should be combined** for predictive purposes.

$\Rightarrow$ **Improved predictions** and **reduced costs and efforts** in the field

- **mRNA and miRNA** were most common in the **optimal** combinations, but for some data sets neither was included.

References – Thank you for your attention!

Boulesteix, A.L., De Bin, R., Jiang, X., Fuchs, M., 2017.

IPF-LASSO: Integrative L1-penalized regression with penalty factors for prediction based on multi-omics data.

Comput. Math. Methods Med., 7691937.

Boulesteix, A.-L., Sauerbrei, W., 2011.

Added predictive value of high-throughput molecular data to clinical data and its validation.

Brief. Bioinform. 12, 215–229.

Efron, B., 1979.

Bootstrap methods: Another look at the jackknife.

Ann. Stat. 7, 1–26.

Herrmann, M., Probst, P., Hornung, R., Jurinovic, V., Boulesteix, A.-L. , 2021.

Large-scale benchmark study of survival prediction methods using multi-omics data.

Brief. Bioinform. 22, bbaa167.

Hornung, R., Wright, M.N., 2019.

Block Forests: random forests for blocks of clinical and omics covariate data.

BMC Bioinform. 20, 1–17.

Ishwaran, H., Kogalur, U.B., Blackstone, E.H., Lauer, M.S., 2018.

Random survival forests.

BMC Bioinform., 19, 1–14.

Klau, S., Jurinovic, V., Hornung, R., Herold, T., Boulesteix, A.L., 1945.

Priority-Lasso: A simple hierarchical approach to the prediction of clinical outcome using multi-omics data.

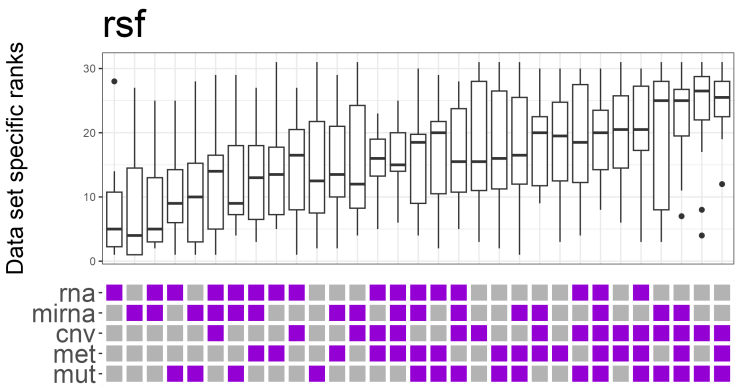
Ann. Math. Stat. 16, 117–186.

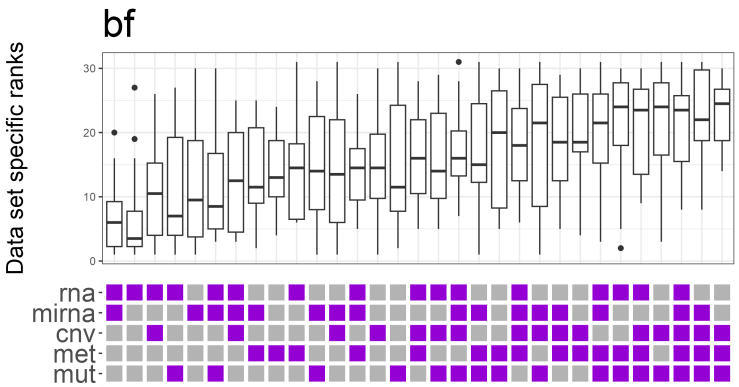
Results: Overview of the considered data sets

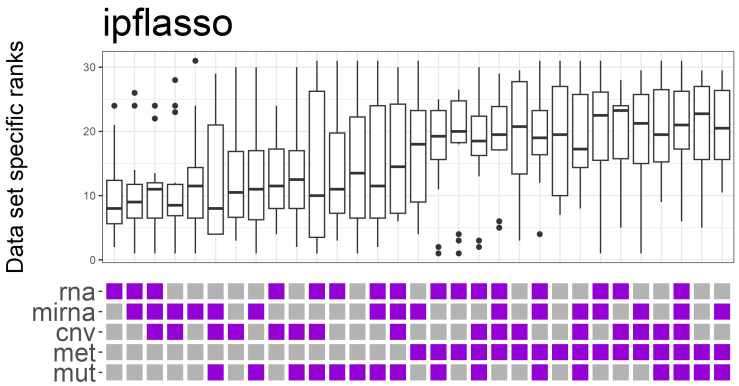
Dataset	Cancer	clin	cnv	mirna	mut	met	rna	f	n	e	r e
BLCA	Bladder urothelial	5	57964	825	18577	382711	23081	483166	382	103	0.27
BRCA	Breast invasive C.	8	57964	835	17975	21919	22694	121398	735	72	0.10
COAD	Colon AC.	7	57964	802	18538	22418	22210	121942	191	17	0.09
ESCA	Esophageal C.	6	57964	763	12628	383295	25494	480153	106	37	0.35
HNSC	Head–neck squamous CC.	11	57964	793	17248	376058	21520	473597	443	152	0.34
LGG	Low grade glioma	10	57964	645	9235	373499	22297	463653	419	77	0.18
LIHC	Liver hepatocellular C.	11	57964	776	11821	378427	20994	469996	159	35	0.22
LUAD	Lung AC.	9	57964	799	18388	22486	23681	123330	426	101	0.24
LUSC	Lung squamous CC.	9	57964	895	18500	21364	23524	122259	418	132	0.32
PAAD	Pancreatic AC.	10	57964	612	12392	375464	22348	468793	124	52	0.42
SARC	Sarcoma	11	57964	778	10001	378139	22842	469738	126	38	0.30
SKCM	Skin cutaneous M.	9	57964	1002	18593	377193	22248	477012	249	62	0.21
STAD	Stomach AC.	7	57964	787	18581	22557	26027	125926	295	62	0.21
UCEC	Uterine corpus EC.	11	57447	866	21053	22517	23978	125875	405	38	0.09

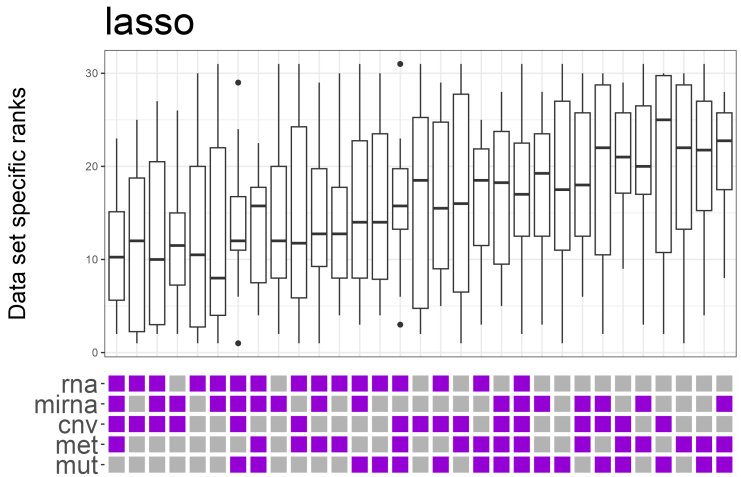
Abbreviations: C. indicates carcinoma; AC., adenocarcinoma; CC., cell carcinoma; M., melanoma; EC., endometrial carcinoma.

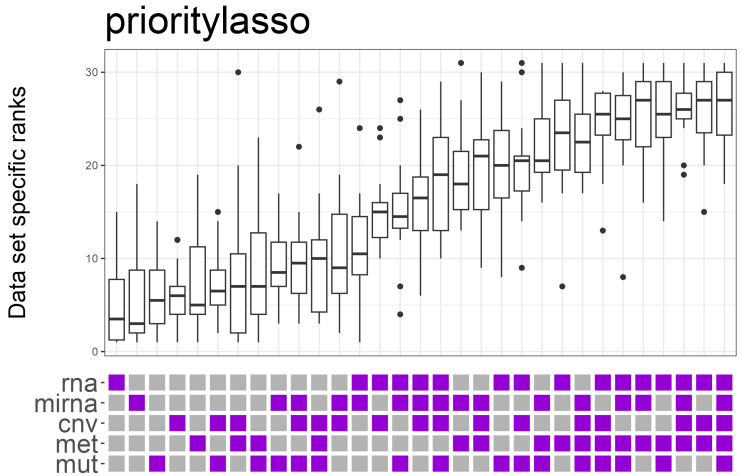
Results: **Ranks** of the data type combinations (`cindex`)



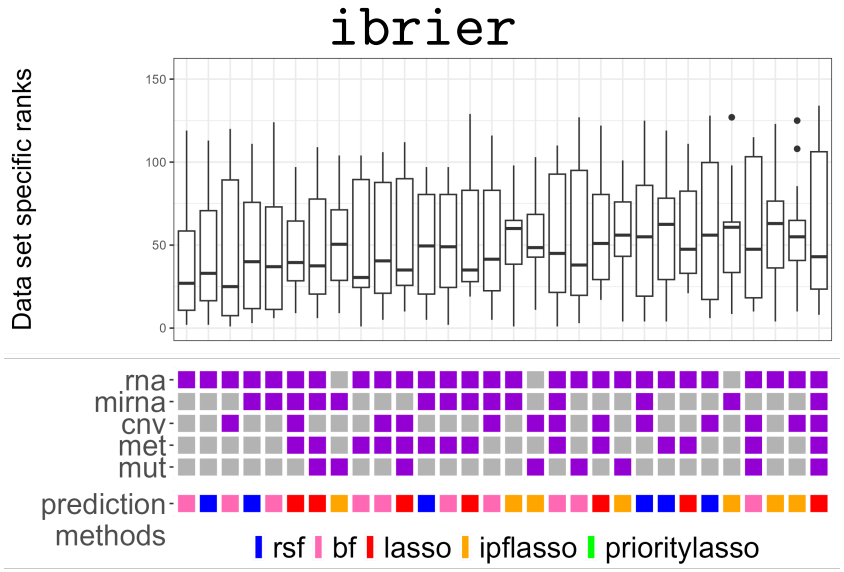


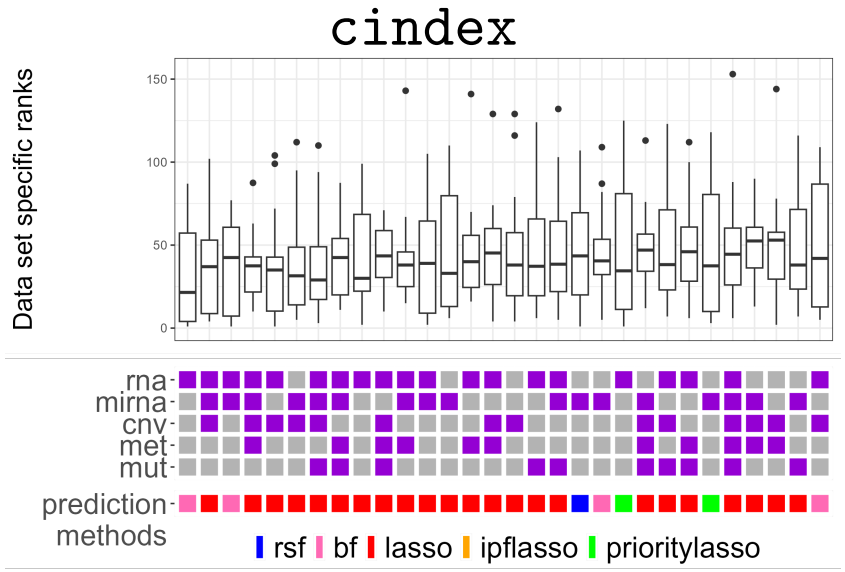




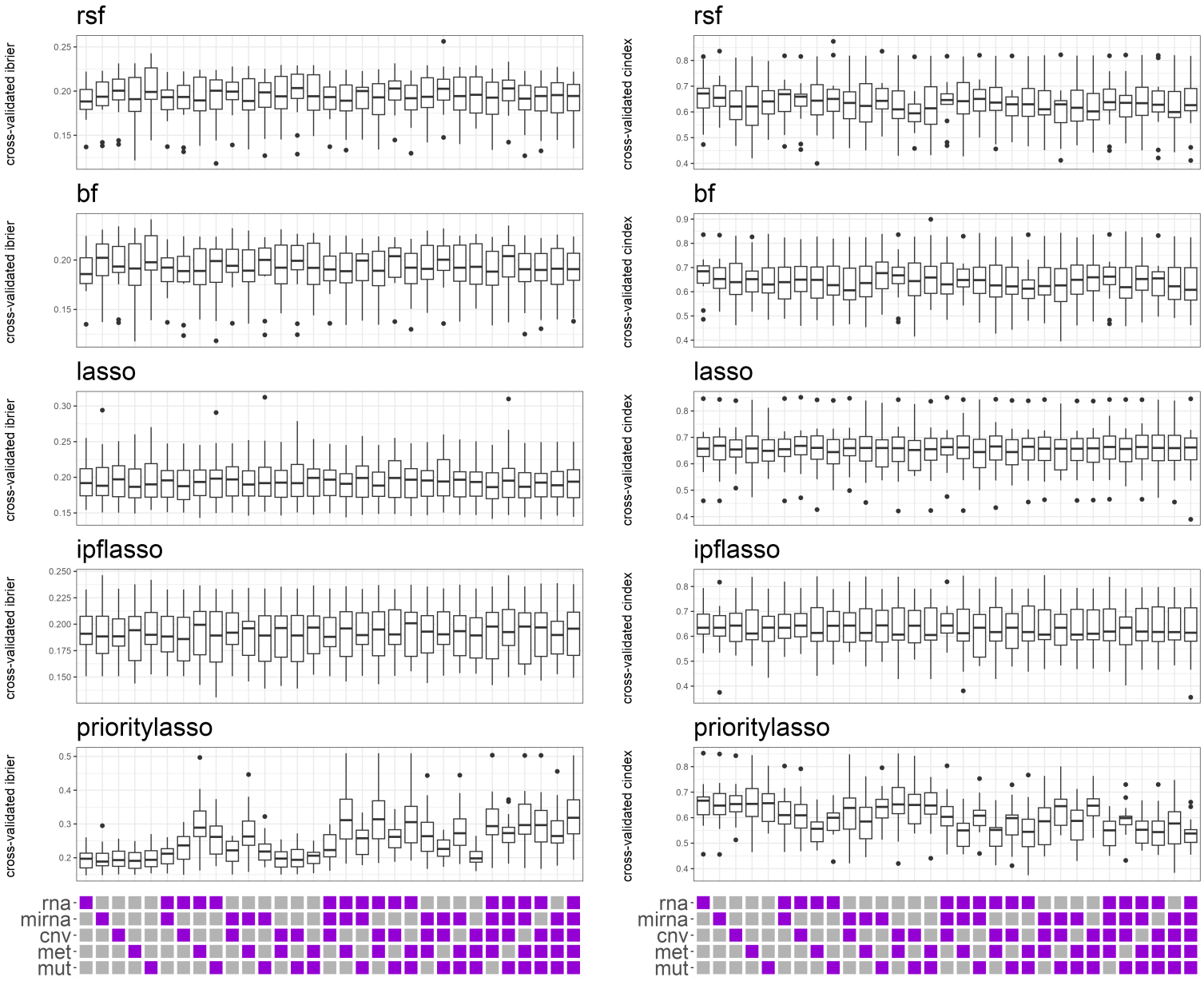


Results: Ranks of each combination of prediction method and data types (30 best combinations)





Results: Cross-validated **performance measure values** for each data type combinations



Results: Mean ranks with 95% percentile intervals obtained using **bootstrap** (`ibrier`) – I

No.	Combination					rsf		bf		ipflasso	
	mut	met	cnv	mirna	rna	mean	ci	mean	ci	mean	ci
1					■	2.8	[1.0, 10.0]	2.0	[1.0, 8.0]	13.4	[2.0, 28.0]
2				■	■	3.2	[1.0, 11.0]	5.1	[1.0, 16.0]	6.8	[1.0, 20.0]
3		■		■	■	3.4	[1.0, 9.0]	5.7	[1.0, 13.0]	14.1	[1.0, 27.0]
4		■			■	5.3	[1.0, 14.0]	5.4	[1.0, 11.0]	21.9	[4.0, 31.0]
5		■	■		■	7.1	[2.0, 14.0]	4.8	[1.0, 10.0]	20.2	[5.0, 30.0]
6			■	■	■	8.4	[3.0, 18.0]	9.7	[4.0, 20.0]	11.6	[1.0, 26.0]
7		■	■	■	■	9.1	[4.0, 16.0]	7.0	[2.0, 14.0]	11.7	[1.0, 25.0]
8			■		■	9.9	[2.0, 20.0]	4.0	[1.0, 12.0]	12.9	[1.0, 29.0]
9	■				■	9.5	[1.0, 19.0]	9.8	[2.0, 21.0]	7.3	[1.0, 22.0]
10	■	■		■	■	9.9	[3.0, 18.0]	12.7	[6.0, 23.0]	17.8	[6.0, 27.0]
11	■	■	■		■	10.8	[3.0, 18.0]	13.2	[7.0, 21.0]	20.6	[6.0, 30.0]
12				■		11.4	[3.0, 22.0]	28.1	[20.0, 31.0]	10.0	[1.0, 27.0]
13	■	■	■	■	■	12.4	[6.0, 18.0]	15.7	[9.0, 24.0]	26.9	[13.0, 31.0]
14	■			■	■	13.0	[5.0, 21.0]	13.1	[5.0, 24.0]	27.1	[15.0, 31.0]
15	■	■			■	12.9	[4.0, 22.0]	14.6	[5.0, 24.0]	17.5	[6.0, 29.0]
16		■		■		14.2	[5.0, 21.0]	14.8	[6.0, 24.0]	13.6	[1.0, 29.0]
17	■		■	■	■	18.0	[9.0, 26.0]	18.1	[7.0, 29.0]	15.9	[3.0, 29.0]
18		■				18.6	[7.0, 28.0]	16.0	[4.0, 28.0]	14.1	[1.0, 29.0]
19		■	■	■		19.1	[12.0, 25.0]	19.5	[10.0, 28.0]	15.2	[1.0, 30.0]
20	■		■		■	20.7	[13.0, 27.0]	19.4	[9.0, 28.0]	22.2	[6.0, 31.0]
21	■			■		20.8	[10.0, 28.0]	21.4	[9.0, 29.0]	5.0	[1.0, 18.0]
22	■	■	■	■		21.8	[15.0, 26.0]	19.1	[8.0, 29.0]	18.6	[5.0, 29.0]
23	■	■		■		21.9	[14.0, 28.0]	22.9	[14.0, 31.0]	20.2	[4.0, 31.0]
24			■	■		22.6	[14.0, 28.0]	19.7	[6.0, 28.0]	12.1	[3.0, 25.0]
25		■	■			23.2	[17.0, 28.0]	19.2	[8.0, 28.0]	16.6	[2.0, 30.0]
26	■	■				25.4	[18.0, 31.0]	23.4	[13.0, 31.0]	25.7	[13.0, 31.0]
27	■	■	■			27.4	[22.0, 31.0]	25.5	[17.0, 31.0]	15.3	[2.0, 29.0]
28	■					27.1	[18.0, 31.0]	29.1	[19.0, 31.0]	21.3	[4.0, 31.0]
29			■			27.8	[21.0, 31.0]	25.6	[14.0, 30.0]	13.6	[2.0, 28.0]
30	■		■	■		28.6	[22.0, 31.0]	26.8	[16.0, 31.0]	19.8	[5.0, 30.0]
31	■		■			29.9	[25.0, 31.0]	24.4	[12.0, 30.0]	7.3	[2.0, 18.0]

Results: Mean ranks with 95% percentile intervals obtained using **bootstrap** (`ibrier`) – II

No.	Combination					lasso		prioritylasso	
	mut	met	cnv	mirna	rna	mean	ci	mean	ci
1					■	23.9	[12.0, 31.0]	2.9	[1.0, 7.0]
2				■	■	15.8	[6.0, 27.0]	9.3	[7.0, 12.0]
3		■		■	■	5.8	[1.0, 16.0]	29.7	[27.0, 31.0]
4		■			■	11.2	[4.0, 24.0]	22.7	[18.0, 27.0]
5		■	■		■	5.4	[1.0, 15.0]	28.5	[25.0, 31.0]
6			■	■	■	12.8	[5.0, 24.0]	14.5	[13.0, 16.0]
7		■	■	■	■	2.4	[1.0, 6.0]	23.9	[18.0, 27.0]
8			■		■	8.2	[1.0, 19.0]	12.8	[10.0, 15.0]
9	■				■	17.4	[6.0, 28.0]	21.1	[16.0, 29.0]
10	■	■		■	■	2.6	[1.0, 7.0]	25.9	[19.0, 30.0]
11	■	■	■		■	6.0	[2.0, 13.0]	23.9	[19.0, 27.0]
12				■		19.8	[8.0, 31.0]	2.9	[1.0, 8.0]
13	■	■	■	■	■	9.2	[3.0, 20.0]	30.1	[27.0, 31.0]
14	■			■	■	8.1	[1.0, 18.0]	27.7	[23.0, 31.0]
15	■	■			■	19.9	[8.0, 30.0]	17.9	[15.0, 24.0]
16		■		■		22.6	[10.0, 31.0]	18.5	[16.0, 24.0]
17	■		■	■	■	10.7	[1.0, 25.0]	25.3	[20.0, 31.0]
18		■				24.3	[11.0, 31.0]	2.8	[1.0, 6.0]
19		■	■	■		20.9	[11.0, 30.0]	16.8	[14.0, 20.0]
20	■		■		■	21.5	[9.0, 31.0]	22.2	[17.0, 29.0]
21	■			■		20.5	[9.0, 29.0]	12.3	[9.0, 16.0]
22	■	■	■	■		8.2	[3.0, 16.0]	20.1	[17.0, 25.0]
23	■	■		■		21.4	[13.0, 30.0]	21.3	[18.0, 25.0]
24			■	■		18.0	[8.0, 28.0]	11.0	[8.0, 14.0]
25		■	■			19.2	[9.0, 30.0]	5.1	[2.0, 8.0]
26	■	■				21.5	[11.0, 31.0]	7.0	[4.0, 10.0]
27	■	■	■			16.6	[7.0, 28.0]	8.9	[5.0, 12.0]
28	■					27.7	[16.0, 31.0]	7.8	[4.0, 11.0]
29			■			23.7	[12.0, 31.0]	2.8	[1.0, 5.0]
30	■		■	■		25.8	[16.0, 31.0]	13.9	[12.0, 16.0]
31	■		■			24.9	[13.0, 31.0]	6.1	[3.0, 9.0]

Results: Mean ranks with 95% percentile intervals obtained using **bootstrap** (`cindex`) – I

No.	Combination					rsf		bf		ipflasso	
	mut	met	cnv	mirna	rna	mean	ci	mean	ci	mean	ci
1					■	2.0	[1.0, 5.0]	1.8	[1.0, 5.0]	3.0	[1.0, 9.0]
2				■		2.6	[1.0, 8.0]	6.8	[3.0, 17.0]	8.1	[1.0, 18.0]
3				■	■	2.8	[1.0, 6.0]	1.6	[1.0, 3.0]	4.0	[1.0, 12.0]
4	■				■	5.7	[1.0, 14.0]	5.8	[1.0, 15.0]	12.8	[4.0, 28.0]
5	■			■		5.9	[2.0, 14.0]	11.8	[4.0, 23.0]	8.3	[1.0, 20.0]
6			■	■	■	8.5	[4.0, 20.0]	9.5	[4.0, 19.0]	4.0	[1.0, 11.0]
7	■			■	■	8.9	[4.0, 20.0]	7.5	[3.0, 18.0]	15.0	[5.0, 30.0]
8		■		■	■	8.8	[3.0, 18.0]	12.7	[6.0, 21.0]	23.3	[7.0, 31.0]
9		■			■	11.4	[5.0, 22.0]	10.6	[4.0, 20.0]	18.2	[4.0, 27.0]
10			■		■	12.3	[5.0, 23.0]	5.2	[2.0, 12.0]	9.5	[5.0, 17.0]
11	■					12.3	[5.0, 22.0]	10.1	[3.0, 19.0]	14.8	[3.0, 23.0]
12		■		■		13.1	[5.0, 26.0]	13.8	[4.0, 26.0]	13.1	[2.0, 29.0]
13			■	■		14.3	[5.0, 26.0]	12.0	[4.0, 24.0]	4.8	[1.0, 13.0]
14		■	■		■	14.4	[7.0, 21.0]	15.4	[6.0, 23.0]	24.2	[13.0, 31.0]
15		■	■	■	■	15.4	[9.0, 22.0]	20.5	[12.0, 27.0]	21.0	[10.0, 30.0]
16	■	■		■	■	15.8	[6.0, 25.0]	24.0	[15.0, 30.0]	22.3	[14.0, 30.0]
17	■	■			■	17.4	[7.0, 26.0]	24.9	[13.0, 31.0]	18.5	[7.0, 29.0]
18	■		■	■	■	17.6	[8.0, 26.0]	18.6	[11.0, 27.0]	17.9	[8.0, 31.0]
19			■			18.9	[7.0, 29.0]	13.8	[5.0, 25.0]	8.6	[4.0, 20.0]
20	■	■				18.7	[8.0, 28.0]	20.4	[10.0, 29.0]	26.7	[16.0, 31.0]
21	■	■		■		20.0	[9.0, 28.0]	19.0	[9.0, 27.0]	27.5	[20.0, 31.0]
22		■	■	■		21.1	[12.0, 27.0]	21.0	[12.0, 28.0]	24.1	[12.0, 31.0]
23		■				21.5	[11.0, 30.0]	10.3	[4.0, 18.0]	21.6	[9.0, 30.0]
24	■		■		■	21.9	[10.0, 29.0]	16.4	[7.0, 27.0]	11.2	[1.0, 28.0]
25	■	■	■	■	■	23.4	[16.0, 28.0]	26.5	[18.0, 31.0]	26.7	[17.0, 31.0]
26		■	■			24.5	[16.0, 29.0]	24.4	[16.0, 30.0]	21.0	[7.0, 30.0]
27	■	■	■		■	24.7	[15.0, 29.0]	25.8	[18.0, 31.0]	18.7	[5.0, 28.0]
28	■		■	■		23.9	[10.0, 30.0]	21.2	[10.0, 30.0]	7.5	[1.0, 23.0]
29	■	■	■	■		28.5	[24.0, 30.0]	29.0	[24.0, 31.0]	23.3	[14.0, 31.0]
30	■		■			29.1	[22.0, 31.0]	25.9	[15.0, 31.0]	11.5	[4.0, 25.0]
31	■	■	■			30.5	[29.0, 31.0]	29.7	[26.0, 31.0]	24.9	[15.0, 31.0]

Results: Mean ranks with 95% percentile intervals obtained using **bootstrap** (`cindex`) – II

No.	Combination					lasso		prioritylasso	
	mut	met	cnv	mirna	rna	mean	ci	mean	ci
1					■	8.2	[1.0, 24.0,]	2.2	[1.0, 6.0,]
2				■		10.3	[1.0, 24.0,]	2.9	[1.0, 8.0,]
3				■	■	9.1	[1.0, 26.0,]	11.5	[6.0, 13.0,]
4	■				■	15.6	[5.0, 28.0,]	20.0	[17.0, 25.0,]
5	■			■		19.3	[7.0, 28.0,]	9.4	[6.0, 13.0,]
6			■	■	■	5.6	[1.0, 18.0,]	15.6	[13.0, 18.0,]
7	■			■	■	14.0	[5.0, 26.0,]	15.0	[13.0, 18.0,]
8		■		■	■	12.1	[3.0, 24.0,]	27.5	[23.0, 31.0,]
9		■			■	12.1	[2.0, 25.0,]	23.0	[19.0, 27.0,]
10			■		■	4.8	[1.0, 16.0,]	14.6	[13.0, 17.0,]
11	■					25.9	[14.0, 31.0,]	18.5	[15.0, 22.0,]
12		■		■		21.0	[7.0, 30.0,]	3.5	[1.0, 8.0,]
13			■	■		6.3	[1.0, 16.0,]	10.9	[6.0, 14.0,]
14		■	■		■	12.2	[2.0, 27.0,]	29.0	[25.0, 31.0,]
15		■	■	■	■	3.9	[1.0, 12.0,]	28.9	[25.0, 31.0,]
16	■	■		■	■	9.1	[2.0, 18.0,]	25.9	[22.0, 30.0,]
17	■	■			■	17.9	[8.0, 27.0,]	27.9	[24.0, 31.0,]
18	■		■	■	■	8.5	[1.0, 20.0,]	18.2	[15.0, 22.0,]
19			■			16.2	[2.0, 29.0,]	3.8	[1.0, 8.0,]
20	■	■				26.1	[14.0, 31.0,]	8.8	[3.0, 13.0,]
21	■	■		■		27.4	[19.0, 31.0,]	22.6	[19.0, 27.0,]
22		■	■	■		22.5	[12.0, 30.0,]	19.4	[16.0, 22.0,]
23		■				25.9	[14.0, 31.0,]	5.4	[1.0, 10.0,]
24	■		■		■	16.8	[6.0, 28.0,]	20.0	[17.0, 24.0,]
25	■	■	■	■	■	18.8	[8.0, 29.0,]	29.4	[25.0, 31.0,]
26		■	■			16.9	[2.0, 30.0,]	7.1	[2.0, 13.0,]
27	■	■	■		■	15.5	[5.0, 25.0,]	25.4	[22.0, 28.0,]
28	■		■	■		24.1	[10.0, 31.0,]	9.7	[6.0, 13.0,]
29	■	■	■	■		18.1	[6.0, 27.0,]	23.9	[21.0, 29.0,]
30	■		■			25.7	[12.0, 31.0,]	5.9	[2.0, 10.0,]
31	■	■	■			26.2	[18.0, 31.0,]	10.3	[5.0, 14.0,]