Aus dem Institut für Medizinische Informationsverarbeitung Biometrie und Epidemiologie (IBE)

Institut der Ludwig-Maximilians-Universität München

Effective use of multi-omics data for prediction of cancer outcomes

Dissertation

zum Erwerb des Doktorgrades der Humanbiologie

an der Medizinischen Fakultät der

Ludwig-Maximilians-Universität München

vorgelegt von

Yingxia Li

aus

Zhoukou, Henan

Jahr

2023

Mit Genehmigung der Medizinischen Fakultät der

Ludwig-Maximilians-Universität zu München

Erster Gutachter:	Prof. Dr. Ulrich Mansmann
Zweiter Gutachter:	Prof. Dr. Tobias Herold
Dritter Gutachter:	Prof. Dr. Axel Imhof

Mitbetreuung durch den promovierten Mitarbeiter:	Dr. Roman Hornung
Dekan:	Prof. Dr. med. Thomas Gudermann

Tag der mündlichen Prüfung: 28.06.2023

Affidavit

Affidavit

Li Yingxia

Surname, first name

Marchioninistr. 15/K U1 832

Street

81377, Munich, German

Zip code, town, country

I hereby declare, that the submitted thesis entitled:

Effective use of multi-omics data for prediction of cancer outcomes

is my own work. I have only used the sources indicated and have not made unauthorised use of services of a third party. Where the work of others has been quoted or reproduced, the source is always given.

I further declare that the submitted thesis or parts thereof have not been presented as part of an examination degree to any other university.

Munich, 28-06-2023 Yingxia Li

place, date *Signature doctoral candidate*

Table of content

List of abbreviations

- **TCGA:** The Cancer Genome Atlas
- **GDC:** Genomic Data Commons
- **CNV:** copy number variation
- **PCA:** Principal Component Analysis
- **LDA:** Linear Discriminant Analysis
- **LUAD:** Lung adenocarcinoma
- **MKL:** multiple kernel learning
- **AUC:** the area under the receiver operating characteristic curve
- **iBrier:** integrated Brier score
- **infor:** information gain
- **mRMR:** The Minimum Redundancy Maximum Relevance
- **Rfe:** recursive feature elimination
- **GA:** genetic algorithm
- **Lasso:** the least absolute shrinkage and selection operator
- **IPF-Lasso:** the integrative Lasso with penalty factors
- **RF-VI:** the permutation importance of random forests

List of publications

Li, Y.; Mansmann, U.; Du, S.; Hornung, R. Synergistic Effects of Different Levels of Genomic Data for the Staging of Lung Adenocarcinoma: An Illustrative Study.Genes. 2021, 12(12), 1872.

Li, Y.; Mansmann, U.; Du, S.; Hornung, R. Benchmark study of feature selection strategies for multi-omics data. BMC Bioinformatics. 2022, 23, 412.

Contribution to the publications

1.1 Contribution to paper I

I designed the experiment, conducted statistical analysis of the data, and wrote the manuscript.

1.2 Contribution to paper II

R.H. and I designed the experiment. I conducted statistical analysis of the data and wrote the manuscript.

1.3 Contribution to unpublished manuscript (Apendix)

R.H. and I designed the experiment. I conducted statistical analysis of the data and wrote the manuscript.

2. Introduction and motivation

With the development of modern biotechnology, different types of high-dimensional molecular data have been generated by high-throughput experiments. This type of data is commonly termed as "omics" data because it includes for example genomics, transcriptomics, proteomics, epigenomics, or metabolomics, referring to the study of the whole genome, the whole transcriptome, the whole proteome, the whole epigenome, or the whole metabolome, respectively. The inclusion of various categories of omics data is often termed multi-omics data.

Multi-omics data have been successfully implemented in numerous applications, such as identifying biomarkers [1], recognizing abnormal pathways in cancer [2] and improving the predictive performance of cancer prognosis and response to treatment [3]. A search on PubMed for the phrase "multi-omics" returned 2 papers in 2006 (first mention of the phrase), 18 papers in 2012, but 1691 papers in 2021. Similarly, a search on PubMed for the phrase "multi-omics and prediction" returned 1 paper in 2008 (first mention of multi-omics and prediction), 5 papers in 2012, but 417 papers in 2021. It can be seen in Figure 2.1 that studying multi-omics has become an increasingly popular topic. In recent years, there has been a growing interest in using multi-omics data to construct predictive models due to the growing abundance of multi-omics data.

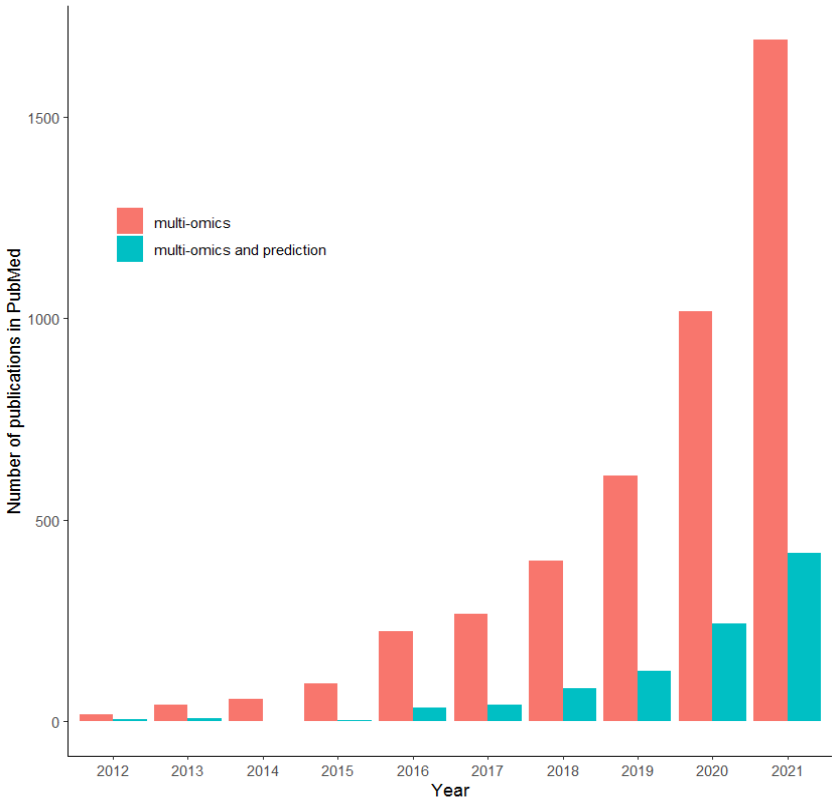

Figure 2.1 Number of results for the search ‘multi-omics’ and ‘multi-omics and prediction’ on Pubmed.

Omics data have been used to develop predictive models for more than 20 years [4]. Early cancer omics studies focused on predictive models generated from single-omics data type or a few different types of omics data simultaneously. For example, Chu et al. used mRNA data to predict relapse and prognosis of colorectal cancer [5]. Patnaik et al. predicted the recurrence of non-small cell lung cancer by miRNA analysis [6]. Dong et al. used DNA methylation data for cancer subtype prediction [7]. Gade et al. used mRNA and miRNA data to improve the predictive performance of clinical outcomes [8]. With the growing availability of other omics data types, the focus has shifted towards constructing predictive models based on multi-omics data, that is, several omics types available for the same patients. For instance, Dong et al. used copy number variation (CNV), mRNA, and methylation data for staging prediction of lung adenocarcinoma (LUAD) [9]. Kim et al. used four types of multi-omics data, including methylation, mRNA, CNV, and miRNA, to predict clinical outcomes [10]. However, few studies have used multi-omics data from the genome to the proteome to demonstrate the role of different types of omics data in the prediction process.

Efficient utilization of multi-omics data has always been a challenge due to the high-dimensional and heterogeneous characteristics. To address the high dimensionality of multi-omics data, feature selection has become a vital component of building predictive models using multi-omics data. The purpose of the feature selection is to select the relevant features and delete the irrelevant ones. Successful feature selection can result in a reduced number of features, shorter training time, and more generalized models. When using multi-omics data for prediction, the most straightforward approach is to combine all datasets, select a subset of features, and then train a learner, which ignores the source of the variables (Figure 2.2A). Instead, as shown in Figure 2.2B, other researchers suggest analyzing each omics data separately and then combining the results [3]. The combination can be done at different stages of the analysis [11]. Currently, a wealth of feature selection approaches exists, and the choice of feature selection methods varies considerably in different literatures. In order to find appropriate feature selection method for omics data, several researches have compared existing methods. For example, Verónica et al. [12] compared 11 feature selection methods, which included 2 embedded methods, 7 filter methods, and 2 wrapper methods, yet the work was performed based on synthetic data. Liu et al. [14] compared 6 feature selection methods using only 2 datasets (Ovarian cancer and Leukemia). Abusamra [13] analyzed 8 different filter-based feature selection methods, which used only mRNA data from Glioma. Although many studies have compared existing feature selection methods [15–17], these studies were often limited in

scope and have not been subjected to large-scale systematic comparisons in the setting of multi-omics data.

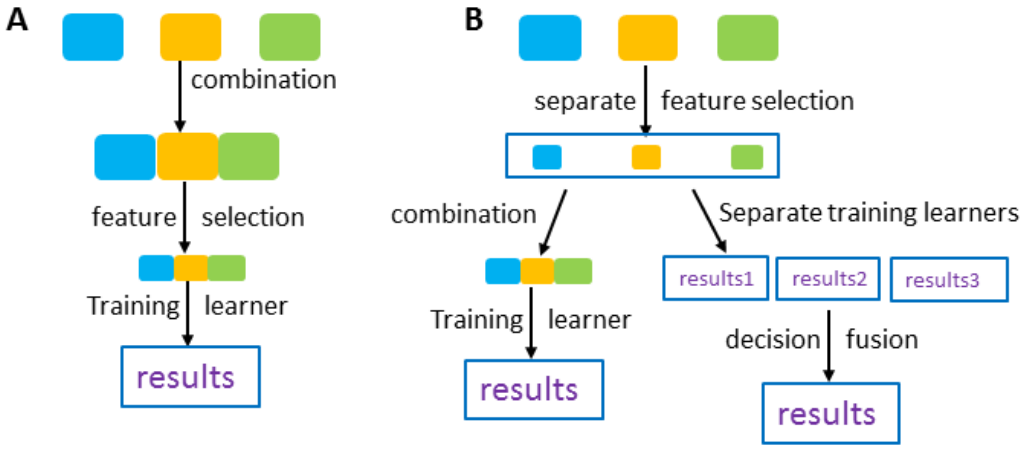

Figure 2.2 A common format of multi-omics data for developing prediction models. Different colored blocks represent different types of omics data.

Although the prediction using multi-omics data has been a well- researched topic, it has to date not been examined, which combinations of omics data tend to work well and which types of omics data correspondingly can, in general, be left out without hurting predictive performance. In fact, it is quite costly and laborious to obtain different omics data types for the same patient. The huge data volume of multi-omics data can also lead to long computation times and large consumptions of computational memory. Moreover, given the costs and efforts needed to obtain multi-omics data, when including multiple omics data types at the same time the sample sizes associated with these data can be expected to be small. Thus, ideally, the number of types of omics data should be small to reduce labor and cost. Several studies have compared the predictive performance of different types of omics data combinations; however, these studies tend to be limited in scope and yielded inconsistent results. For instance, a pioneering study by Zhao et al. [3] conducted a comparative study considering four types of genomic data. They used four datasets and prediction methods that do not consider the multi-omics structure (more sophisticated methods were not yet available at the time of conduction of their analysis). It was observed in that study that once mRNA data and clinical covariates were included in the model, the addition of any further genomic data types did not substantially improve prediction results. On the other hand, a similar study [18] considered four types of genomic data, each measured for different patients. They performed predictive modelling using each type of omics data individually and concurrently using all types of omics data taken together. Here, first, the integration of four types of omics data produced a slightly better model performance than using any of the single omics data, and second, for the

individual omics data, the best predictions were obtained using mRNA, followed by miRNA and CNV. It is, however, unclear, which combinations of omics data tend to improve the predictive performance and which types of omics data can correspondingly, in general, be left out to save efforts and costs.

In this thesis, we aim to find appropriate methods and strategies for the effective use of multi-omics data to achieve optimal prediction results. This thesis includes three different papers that have been or will be published in scientific journals. For ease of reading, I refer to these papers as contribution 1, contribution 2, and contribution 3 below, where contribution 1 and contribution 2 refer to the published papers of Li et al. (2021) and Li et al. (2022a), respectively, and contribution 3 refers to the unpublished paper of Li et al. (2022b). Contribution 1 used a self-developed method to demonstrate the role of different types of omics data from the genome to the proteome in the context of LUAD stage prediction. In contribution 2, given that multi-omics data have a special structure that varies from single-omics data, we conducted a large-scale benchmarking study to compare different feature selection strategies and methods for multi-omics data. Specifically, we used 15 cancer datasets with binary outcomes and compared eight commonly used feature selection methods under several scenarios that may affect the prediction performance. In contribution 3, by aims of a large-scale benchmark experiment we aim to address whether all types of omics data are necessary for prediction purposes or whether, in contrast, some types of omics data can be left out without hurting the predictive performance notably. We compared the predictive performance of all possible 31 combinations of 5 types of genomic data (mRNA, miRNA, methylation, mutation, and CNV) using 14 cancer datasets with survival outcome.

In the next section of this thesis, I will present the concepts and methods of this work in more detail. In Section 2.1, I will introduce the multi-omics data and their interactions. In Section 2.2, feature selection methods will be introduced. In Section 2.3, the prediction methods for multi-omics data will be described.

2.1 Overview of multi-omics data

Data that contain several types of omics measurements are termed multi-omics data. These usually refer to such as epigenomics, genomics, proteomics, transcriptomics, and metabolomics.

Genomics is the study of the structure and function of the whole genome, that is, the entire DNA sequence. Genomic analysis focuses on mutations, insertions, deletions, CNV, and structural variation [19]. whole-genome sequencing and

Whole-exome sequencing are two popular technologies used in genomic studies [20].

Epigenetics refers to the hereditary changes in gene function caused by biochemical modifications [21]. It includes DNA methylation, chromatin remodeling, histone modifications, and ribonucleic acid inference, which act together to regulate gene transcription and maintain genomic stability [22,23]. DNA methylation can be measured by a variety of methods, such as bisulfite sequencing [24]. High throughput measurement of histone modifications can be done with chromatin immunoprecipitation sequencing experiments [25].

Transcriptomics is the term given to the qualitative and quantitative study of complete sets of transcripts, including coding and non-coding [26]. It is highly dynamic in the human transcriptome, varying strongly between cellular states [27] and tissues [28], and showing short-lived responses to environmental stimuli, such as dietary changes [29,30]. There are several methods available to quantify the transcriptome, the most popular being RNA sequencing and microarray [31], the former now being more commonly used because it provides better performance and data consistency [32].

Proteomics is the investigation of proteins and their interactions [33]. To some extent, the proteome represents the underlying transcriptome. The proteome is also not constant, it varies from cell to cell and changes over time. High throughput measurement of proteomics can be done with mass spectrometry [34].

Metabolomics is the comprehensive study of all metabolites within a cell, tissue, or organism under a given set of conditions [30,35,36]. Notably, metabolomics data are not saved in a gene-level format, which makes it difficult to combine metabolomics data with other histological data to predict clinical outcomes.

2.1.1 Multi-omics data interaction

Genomics, epigenomics, proteomics, transcriptomics, and metabolomics interact with each other to maintain the normal function of this biological system, as shown in Figure 2.3. If a system is disturbed by environmental influences or pathophysiological processes, several or even all system levels are involved. Different omics data provide complementary information for understanding the process of disease onset [30,37]. For example, mutations and methylation of DNA may lead to alterations in transcription, translation, post-translational modifications, and ultimately gene and protein function. Published studies have illustrated that multi-omics data are important for understanding cancer biology [38,39], as single-omics data cannot provide all the information in the constitutive mechanisms. It is, however, unclear, which combinations of omics data tend to work well and which

types of omics data correspondingly can, in general, be left out without hurting predictive performance.

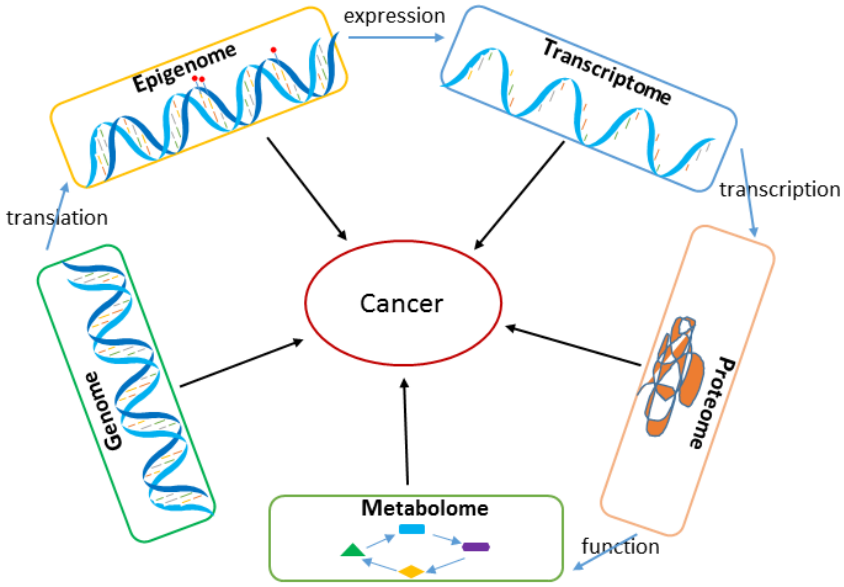

Figure 2.3 Multiple layers of omics data in biological system.

In contribution 1, we illustrated the role of different omics data types from the genome to the proteome in the context of LUAD stage prediction. In contribution 3, using a large-scale benchmark experiment, we investigate whether all types of omics data are necessary for prediction purposes or whether, in contrast, some types of omics data can be left out without hurting the predictive performance notably. We used 14 cancer datasets and 5 prediction methods to compare the predictive performance of survival outcomes for all 31 possible combinations of 5 types of omics data from genomics, transcriptomics, and epigenetics.

2.1.2 TCGA dataset

The Cancer Genome Atlas (TCGA, https://tcga-data.nci.nih.gov/tcga/) [40] project is a joint effort started in 2006 by the National Human Genome Research Institute and the National Cancer Institute. It has generated multi-omics data on over 20,000 primary cancer samples from 33 cancer types. Raw sequencing data, transcriptome profiling, proteome profiling, copy number variant, single nucleotide variant, methylation information, clinical information, and biospecimen supplement are available in the project. All of the over 2.5 petabytes of data generated through TCGA are currently available at the Genomic Data Commons (GDC, https://portal.gdc.cancer.gov/), including clinical, molecular and imaging data.

In this thesis, all multi-omics data and clinical information used are both from TCGA.

2.2 Feature selection for multi-omics data

A dataset with observations n much smaller than the variable p, that is, n<<p, is called high-dimensional data. Most multi-omics datasets have the classic "curse of dimensionality" problem, usually with more than 10 000 or even 100 000 variables, that is, multi-omics features (p) are much larger than the observation sample (n) [41]. These high-dimensional multi-omics data often contain many redundant, uncorrelated, or weakly correlated features, which can mislead the training of algorithms [42].

When analyzing high-dimensional data, feature extraction and feature selection are available to reduce the dimensionality or apply methods that consider high dimensions directly. Feature extraction is the construction of a new feature subspace using some algorithms, which summarizes the original dataset and its dimensionality, such as Principal Component Analysis [43] and Linear Discriminant Analysis [44]. Feature selection is the removal irrelevant or redundant features by some algorithms to obtain a subset from the original features. Feature selection techniques can pre-process learning algorithms, reduce learning time, improve learning performance, and simplify learning outcomes [45–47]. Both theory and practice have demonstrated the effectiveness of feature selection in processing high-dimensional data and improving learning efficiency [48,49].

According to different principles, feature selection approaches could be classified into various categories. According to the relationship between feature selection and learning methods, they can be divided into embedded, filter, wrapper, and hybrid methods [50–58], which are the most frequent classifications. Based on the output type, they could be divided into subset evaluation and individual evaluation [59]. In addition, based on the evaluation criterion, they can be derived from dependence, correlation, consistency, information, and Euclidean distance measure [60]. Based on the search strategies, they can be classified into forward selection, backward elimination, random, and hybrid models [50].

So far, several studies have performed comparisons of existing feature selection algorithms, but none have performed comparisons of feature selection methods in the context of the specific structure of multi-omics data. As seen in Figure 2.2, feature selection can be performed by applying to individual omics datasets and then combining them, or by applying directly to the combined omics datasets. As these data have a particular structure that differs from single-omics data, it is not clear whether different feature selection strategies would work well on these datasets.

In contribution 2, after conducting an extensive literature survey, we analyzed 8 commonly used feature selection methods, including 2 wrapper methods, 2 embedded methods, and 4 filter methods. Given the specific structure of multi-omics data, we also made a comparison of the performance of the same methods under different strategies, namely, the number of selected features, feature selection type: selection from all data types simultaneously or separately, and including versus excluding clinical data. In contribution 2, we introduced the concepts of filters, wrappers, and embedded methods in detail. When presenting the results, we divided these methods into subset evaluation and individual evaluation. Therefore, the concepts of subset evaluation and individual evaluation are introduced next.

2.2.1 Individual evaluation

Individual evaluations, also termed feature ranking/weighting, assign weights by evaluating the relevance of individual variables to the target class. Typically, a subset of features is selected from the top of the ranked list. This method is very effective for high-dimensional data because it has a linear time complexity in dimensionality. However, this evaluation approach does not take into account the correlation between features, that is it does not remove redundant features, which may have similar rankings. Once features are considered relevant to the target class, they are all selected, even if many of them are highly relevant. For high-dimensional data that may contain a great number of redundant features, this method may produce results that are far from optimal. However, it is also frequently selected due to its low time expenditure.

In contribution 2, the permutation importance of random forests (RF-VI), reliefF, information gain (infor), the minimum redundancy maximum relevance (mRMR), and t-test are all individual evaluation methods.

2.2.2 Subset evaluation

Subset evaluation generates a candidate subset of features based on a certain search strategy. Each candidate subset is evaluated by a certain evaluation metric and compared to the previous best subset. If a new subset proves to be better, it will replace the previous best subset. The process of subset generation and evaluation is repeated until a given stopping criterion is met. Unlike individual evaluations, the evaluation measures used in this method are defined for a subset of features, taking into account the presence and influence of redundant features. This method selects a subset of features that is close to the optimal subset. However, in the subset generation step, a search of the feature subset is required, and although various heuristic search strategies exist, such as greedy

sequential search, best-first search, and genetic algorithms, most of them still incur an n^2 time complexity (n is the number of features), which prevents them from scaling well to datasets containing tens of thousands of features.

In contribution 2, genetic algorithm (GA), recursive feature elimination (Rfe), and the least absolute shrinkage and selection operator (Lasso) are all subset evaluation methods.

2.3 Prediction methods for multi-omics data

Omics data have been used to develop predictive models for more than 20 years [4]. Early cancer omics studies focused on predictive models generated from single-omics data type. With the growing availability of other omics data types, the focus has shifted towards constructing predictive models based on multi-omics data, that is, several omics types available for the same patients. Basically, in two ways, multi-omics data can be incorporated into predictive models, that is, whether to consider the group structure of multi-omics data or not.

A total of 8 prediction methods were used in our work. In contribution 1 we used multiple kernel learning (MKL), random forest, K-nearest neighbors, logistic regression, and support vector machine, for the prediction of LUAD stages. In contribution 2 we used random forest and support vector machine for a binary prediction. In contribution 3 we used random forest, block forest, the least absolute shrinkage and selection operator (Lasso), the two-step integrative lasso with penalty factors (IPF-Lasso), and priority-Lasso to predict survival outcomes. Among these methods, MKL, block forest, IPF-Lasso, and priority-Lasso consider multi-omics data structures.

MKL considers the group structure of multi-omics data by applying different kernels to different omics data types.

The random forest algorithm, a tree-based ensemble approach first proposed by Breiman [61], has now grown to a standard non-parametric classification and regression tool that is used in many different fields. A random forest consists of several hundred to several thousand decision trees. Block forest [62] is a variant that modifies the random forests split point selection to incorporate the group structure of multi-omics data. In block forests, it calls each omics type a block, which is the origin of the block forest name.

Standard Lasso [63] is a penalized regression method that applies L1-regularization to penalize the variables and shrinks some of the coefficients to zero. Thus, the lasso can be used not only for constructing predictive models but also for variable selection. Priority-Lasso [64] is a lasso-based hierarchical regression

method that fits lasso models successively to several data blocks with different priorities and takes the predicted value as the offset for the next data block. The IPF-Lasso [65] is also an extension of the standard lasso that considers the group structure by using different values of penalty parameters for each data type.

3. Summary (in Englisch)

Today, various categories of omics data exist. These include epigenomics, genomics, proteomics, transcriptomics, and metabolomics data. Early cancer omics studies focused on predictive models generated from single-omics data type. As studies on the performance of different types of omics data in predictive models have accumulated, researchers have found that multi-omics data are superior to single-omics data in several ways. However, effective use of multi-omics data for making the predictions is challenging for various reasons. First, the prediction information varies between different omics data types, but at the same time there is an interaction between them. Second, large amounts of omics data are useless for prediction where they are either irrelevant or redundant. Third, which combinations of omics data tend to work well and which types of omics data correspondingly can, in general, be left out without hurting predictive performance.

To address the above issues, we first investigated the role of individual omics data in prediction. Second, to obtain an optimized feature subset, we studied the feature selection methods that are most suitable for multi-omics data through a large-scale experiment. Finally, we investigated through a large-scale benchmark experiment whether all types of omics data in multi-omics data are necessary for prediction purposes.

Contribution 1 can be considered as a practical description to illustrate the potential of different types of omics data in the context of LUAD stage prediction. A self-developed method, Omics-MKL, was used. It integrates an existing feature ranking technique mRMR, which retains the features most relevant to the predicted target class while reducing redundancy between these features, and MKL, which applies different kernels to different types of omics data. We found that each considered omics data type provided helpful prediction information, but differed in their ability to provide predictive information, with mRNA data and methylation data playing important roles in LUAD staging prediction. Moreover, the prediction performance of multi-omics data was superior to the prediction performance of single-omics data.

Contribution 2, we performed a large-scale benchmark study to investigate feature selection methods and strategies for multi-omics data. After an extensive survey of the literature over the past 10 years, we obtained 8 common feature selection approaches for cancer classification, including 4 filter methods (t-test, infor, reliefF, and mRMR), 2 wrapper methods (GA and Rfe), and 2 embedded methods (Lasso and RF-VI). In addition to comparing the feature selection methods, we also compared the performance of the same methods under different strategies, namely, the number of selected features, the type of feature selection:

selection from all data types simultaneously or separately, and including versus excluding clinical data. We used 15 cancer data types from TCGA, where each cancer data contained 4 types of molecular data (mRNA, miRNA, CNV, and mutation) and clinical data with the presence versus absence of TP53 mutation as the outcome variable for classification. We used support vector machine and random forests as classifiers to evaluate the prediction performance with accuracy, AUC, and Brier score, and the feature selection time was also one of our evaluation metrics. We evaluated the performance of each method on each dataset using a 5-fold cross-validation method repeated three times. It was found that, first, for most of the considered feature selection methods, the number of chosen features impacted the performance of the prediction. Second, including or excluding clinical information and the type of feature selection did not have a large impact on prediction performance. Third, the computational cost of the wrapper methods is much higher than that of the embedded and filter methods, and for some methods (mRMR and reliefF) separate selection takes less time. Finally, regardless of the performance metric considered, the feature selection approaches Lasso, mRMR, and RF-VI tended to be superior to the other approaches considered. Here, RF-VI and mRMR already delivered powerful prediction performance when considering only a few selected features. Therefore, we recommend mRMR and RF-VI for feature selection of multi-omics data, where, however, mRMR is much more computationally expensive.

Contribution 3, we conducted a large-scale benchmark study to address whether all types of omics data are necessary for prediction purposes or whether, in contrast, some types of omics data can be left out without hurting the predictive performance notably. We compared the predictive performance of all possible 31 combinations of 5 types of genomic data (mRNA, miRNA, methylation, mutation, and CNV) using 14 cancer datasets with survival outcome. As prediction methods, we used random forests, block forests, Lasso, priority-Lasso, and IPF-Lasso. These methods were evaluated with repeated 5-fold cross validation. The integrated Brier score and Harrell's concordance index were used as performance metrics. To investigate the stability of the results, bootstrap analysis at the level of the included datasets was performed. Contrasting our expectations, we did not generally see an improved predictive performance by combining several omics data types. Instead, using only mRNA data or a combination of mRNA and miRNA data was sufficient in most cases. Combinations of larger numbers of omics data types tended to lead to a worsening of predictive performance. our results indicate that using only few data types tends to be associated with better performance. In most cases mRNA or combinations of mRNA and miRNA are sufficient, but for some datasets also other omics data types are important. As

found previously [66] irrespective of the included omics data types it is important to also include clinical covariates and to prioritize them in the prediction.

In Contribution 1, we combined the feature selection method mRMR with the classifier MKL to illustrate the effect of different levels of omics data on LUAD stage prediction. However, in contribution 2, we did not use MKL as our classifier for two reasons. First, Contribution 2 focused mainly on the feature selection algorithms, and we did not need classifiers that considered multi-omics data group structure. Second, all the work in contribution 1 was run on Matlab, while contribution 2 was run on R. MKL runs fast in Matlab and takes much time in R. It was also not considered as a classifier in contribution 3 due to the running time. In Contribution 2, we found that RF-VI had good performance in multi-omics feature selection, and that simultaneous and separate selection did not have a considerable effect on prediction performance. So in the next contribution 3, we used these findings. For individual omics data with more than 2500 variables, we used the RF-VI feature selection method to select features. In both contributions 1 and 3 we found that mRNA played a particularly important role in improving the predictive performance of clinical outcomes. In contribution 3 we found that the combination of larger numbers of omics data categories did not improve the predictive performance, whereas in contribution 1 we found that the prediction performance of multi-omics data outperformed that of single-omics data. This may be due to the fact that in Contribution 1 we did not include clinical variables and the predicted targets were different.

Our work has some limitations. First, this work considers only internally validated performance estimate, which may make the obtained results over-optimistic. Second, the number of datasets and the sample size are not large enough; there are not many available datasets with over 100 samples and multi-omics data in TCGA. Third, our data sources are not diverse, we only used data from TCGA.

4. Zusammenfassung (deutsch)

In den letzten Jahren sind für individuelle Patienten verschiedene Arten von Omics-Daten wie Genomics, Epigenomics, Transcriptomics, Proteomics und Metabolomics simultan verfügbar geworden. Frühe auf Omics-Daten basierende onkologische Studien konzentrierten sich hauptsächlich auf einen einzelnen Omics-Datentyp oder wenige verschiedene Omics-Datentypen gleichzeitig, um Vorhersagemodelle zu erstellen. Mit zunehmender Erfahrung mit Omics basierten Studien für Prädiktionsmodelle wurde klar, daß Prädiktionsmodelle basierend auf Multi-Omics-Ansätzen in mehreren Aspekten besser abschneiden als solche basierend auf Single-Omics-Daten. Die effektive Nutzung von Multi-Omics-Daten für die Vorhersage ist jedoch aus mehreren Gründen schwierig. Erstens variiert der Grad der prädiktiven Information zwischen den einzelnen Omics-Daten und es gibt Wechselwirkungen zwischen ihnen. Zweitens ist eine große Menge von Omics-Daten für die Vorhersage nicht informativ, da sie entweder redundant oder irrelevant sind. Drittens ist es schwierig zu entscheiden, welche Kombinationen von Omics-Daten gut funktionieren und welche Arten von Omics-Daten dementsprechend im Modell weggelassen werden können, ohne die Vorhersageleistung zu beeinträchtigen.

Um die oben genannten Fragen zu klären, untersuchten wir zunächst die Rolle der einzelnen Omics-Daten bei der Vorhersage. Um die Dimensionalität der Multi-Omics-Daten zu reduzieren, haben wir Methoden zur Merkmalsauswahl in groß angelegten Simulationsexperimenten untersucht. Schließlich untersuchten wir anhand eines groß angelegten Benchmark-Experiments, ob alle Arten von Omics-Daten für die Vorhersage notwendig sind oder ob gewisse Omics-Daten weggelassen werden können, ohne die Vorhersage-Performance merklich zu beeinträchtigen

Beitrag 1 leifert Veranschaulichung veranschaulicht des Potenzials von Multi-Omics-Daten im Zusammenhang mit der Vorhersage des LUAD-Stadiums. In diesem Beitrag verwenden wir eine selbstentwickelte Methode, Omics-MKL, um die Rolle der spezifischen Omics-Datentypen im Vorhersageprozess zu untersuchen.Omics-MKL integriert ein bestehendes Feature-Ranking-Verfahren (mRMR), dass die für die vorhergesagte Zielklasse relevantesten Merkmale beibehält und gleichzeitig die Redundanz unter ihnen reduziert. Wir verwenden MKL, das unterschiedliche Kernel für verschiedene Omics-Datentypen zu ermöglicht. Wir zeigen, dass alle betrachteten Omics-Daten nützliche Vorhersageinformationen liefern, sich aber in ihrer prädiktiven Fähigkeit unterschieden. Dabei spielen mRNA- und Methylierungsdaten eine wichtige Rolle bei der Vorhersage des

LUAD-Stadiums. Außerdem war die Vorhersageleistung von Multi-Omics-Daten besser als die von Single-Omics-Daten.

In Beitrag 2 untersuchen wir mittels einer groß angelegten Benchmark-Studie Methoden und Strategien zur Merkmalsauswahl für Multi-Omics-Daten. Nach einer umfangreichen Literaturrecherche zur Methodenentwicklung in den letzten 10 Jahren erhielten wir acht Verfahren zur Merkmalsauswahl, darunter vier Filtermethoden (t-Test, infor, reliefF und mRMR), zwei Wrapper-Methoden (Rfe und GA) und zwei eingebettete Methoden (Lasso und RF-VI). Zusätzlich zum Vergleich der Merkmalsauswahlmethoden haben wir auch deren Leistung unter verschiedenen Strategien bezüglich der Anzahl der ausgewählten Merkmale, die Art der Merkmalsauswahl (Auswahl aus allen Datentypen gleichzeitig oder getrennt), und Einbeziehung bzw. Ausschluss klinischer Daten untersucht. Wir verwendeten Daten zu 15 Tumortypen aus dem TCGA, wobei jeder Krebsdatentyp vier Arten von molekularen Daten (mRNA, miRNA, CNV und Mutation) und klinische Daten enthielten. Vorherzusagen war, ob eine TP53-Mutation vorliegt oder nicht. Als Klassifikatoren verwenden wir Support Vector Machine (SVM) und Random Forests. Wir untersuchendie Vorhersageleistung über folgende Bewertungsmetriken: Genauigkeit, AUC und Brier-Score. Als Bewertungsschema verwendeten wir eine 5-fache Kreuzvalidierung, die dreimal wiederholt wurde, um die Leistung der einzelnen Methoden für jeden Datensatz zu messen. Wir stellten fest, dass bei den meisten der betrachteten Methoden zur Merkmalsauswahl die Anzahl der ausgewählten Merkmale die Vorhersageleistung beeinflusste. Weiterhin hatten die Einbeziehung oder der Ausschluss klinischer Informationen und die Art der Merkmalsauswahl keinen wesentlichen Einfluss auf die Vorhersageleistung. Drittens sind Wrapper-Methoden wesentlich rechenintensiver als Filter- und eingebettete Methoden, während bei einigen Methoden (mRMR und reliefF) die separate Auswahl weniger Zeit in Anspruch nimmt. Unabhängig von der betrachteten Leistungsmetrik schneiden die Merkmalsauswahl methoden mRMR, RF-VI und Lasso tendenziell besser ab als die anderen betrachteten Methoden. Dabei lieferten mRMR und RF-VI bereits eine starke Vorhersageleistung, wenn nur wenige ausgewählte Merkmale berücksichtigt werden. Daher empfehlen wir RF-VI und die Filtermethode mRMR für die Merkmalsauswahl bei Multi-Omics-Daten, wobei mRMR allerdings deutlich rechenaufwendiger ist.

In Beitrag 3 haben wir eine groß angelegte Benchmark-Studie durchgeführt. Wir untersuchen, ob alle Arten von Omics-Daten für die Vorhersage notwendig sind oder ob im Gegensatz dazu einige Arten von Omics-Daten weggelassen werden können, ohne die Vorhersageleistung nennenswert zu beeinträchtigen. Wir haben die Vorhersageleistung aller 31 möglichen Kombinationen von 5 Arten von Genomdaten (mRNA, miRNA, Methylierung, Mutation und CNV) anhand von 14

onkologischen Datensätzen mit dem Endpunkt OS (overall survival) verglichen. Als Vorhersagemethoden verwendeten wir Random Forrests, Block Forrests, Lasso, Prioritäts-Lasso und IPF-Lasso. Diese Methoden wurden durch wiederholte 5-fache Kreuzvalidierung verglichen. Der integrierte Brier-Score und der Konkordanzindex von Harrell wurden als Leistungskennzahlen verwendet. Um die Stabilität der Ergebnisse zu untersuchen, wurde eine Bootstrap-Analyse auf der Ebene der einbezogenen Datensätze durchgeführt. Im Gegensatz zu unseren Erwartungen konnten wir im Allgemeinen keine verbesserte Vorhersageleistung durch die Kombination mehrerer Omics-Datentypen feststellen. Stattdessen reichte es in den meisten Fällen aus, nur mRNA-Daten oder eine Kombination aus mRNA- und miRNA-Daten zu verwenden. Kombinationen einer größeren Anzahl von Omics-Datentypen führten tendenziell zu einer Verschlechterung der Vorhersageleistung.

Unsere Ergebnisse zeigen, dass die Verwendung von nur wenigen Datentypen tendenziell mit einer besseren Leistung verbunden ist. In den meisten Fällen sind mRNA oder Kombinationen aus mRNA und miRNA ausreichend, aber für einige Datensätze sind auch andere Omics-Datentypen wichtig. Wie bereits festgestellt [68], ist es unabhängig von den einbezogenen Omics-Datentypen wichtig, auch klinische Kovariaten einzubeziehen und sie bei der Vorhersage zu priorisieren.

In Beitrag 1 haben wir die Merkmalsauswahlmethode mRMR mit dem Klassifikator MKL kombiniert, um die Auswirkungen verschiedener Ebenen von Omics-Daten auf die Vorhersage des LUAD-Stadiums zu veranschaulichen. In Beitrag 2 haben wir jedoch aus zwei Gründen nicht MKL als Klassifikator verwendet. Erstens konzentrierte sich Beitrag 2 hauptsächlich auf die weit genutzten Algorithmen zur Merkmalsauswahl, und wir brauchten keine Klassifikatoren, die die Struktur von Multi-Omics-Data-Gruppen berücksichtigen. Zweitens wurde die gesamte Arbeit in Beitrag 1 in Matlab ausgeführt, während Beitrag 2 in R ausgeführt wurde. MKL läuft schnell in Matlab und benötigt viel Zeit in R. Aufgrund der Laufzeit wurde es auch nicht als Klassifikator in Beitrag 3 berücksichtigt. In Beitrag 2 stellten wir fest, dass RF-VI bei der Auswahl von Multi-Omics-Features eine gute Leistung erbrachte und dass die gleichzeitige und getrennte Auswahl keinen wesentlichen Einfluss auf die Vorhersageleistung hatte. Im nächsten Beitrag 3 haben wir diese Ergebnisse verwendet. Für individuelle Omics-Daten mit mehr als 2500 Variablen verwendeten wir die RF-VI-Merkmalsauswahlmethode, um Merkmale auszuwählen. Sowohl in Beitrag 1 als auch in Beitrag 3 stellten wir fest, dass die mRNA eine besonders wichtige Rolle bei der Verbesserung der Vorhersageleistung klinischer Ergebnisse spielt. Beitrag 3 macht weiterhin klar, dass die Kombination einer größeren Anzahl von Omics-Datentypen die Vorhersageleis-

tung nicht verbesserte, während wir in Beitrag 1 feststellten, dass die Vorhersageleistung von Multi-Omics-Daten bei einem bestimmten Verfahren besser war als die von Single-Omics-Daten. Dies könnte darauf zurückzuführen sein, dass wir in Beitrag 1 keine klinischen Variablen einbezogen haben und die vorhergesagten Ziele anders waren. Unsere Arbeit hat einige Einschränkungen. Erstens berücksichtigt diese Arbeit nur intern validierte Leistungsschätzungen, was die erzielten Ergebnisse möglicherweise zu optimistisch erscheinen lässt. Zweitens sind der Stichprobenumfang und die Anzahl der Datensätze nicht groß genug; es gibt nicht viele verfügbare Datensätze mit mehr als 100 Proben und Multi-Omics-Daten in TCGA. Drittens, wir verwenden nur TCGA als Datenquelle. Die Messung der Omics-Daten folgt im TCGA harmonisierten Protokollen. Somit sind diese Datensätze nicht typischen Quellen von Heterogenität unterworfen.

5. Paper I

Synergistic Effects of Different Levels of Genomic Data for the Staging of Lung Adenocarcinoma: An Illustrative Study

This paper was published in Genes:

Li, Y., Mansmann, U., Du, S., & Hornung, R. (2021). Synergistic effects of different levels of genomic data for the staging of lung adenocarcinoma: An illustrative study. *Genes*, *12*(12), 1872.

DOI: https://doi.org/10.3390/genes12121872

6. Paper II

Benchmark study of feature selection strategies for multi-omics data

This paper was published in BMC bioinformatics:

Li, Y., Mansmann, U., Du, S., & Hornung, R. (2022). Benchmark study of feature selection strategies for multi-omics data. *BMC bioinformatics*, *23*(1), 1-18.

DOI: https://doi.org/10.1186/s12859-022-04962-x

Apendix A: Unpublished Manuscript

Can combining many data types in multi-omics data lead to a worsening of predictive performance? A large-scale benchmark study

Yingxia Li[1], Ulrich Mansmann[1], Roman Hornung[1]

[1]Institute for Medical Information Processing, Biometry and Epidemiology, University of Munich, Marchioninistr. 15, 81377 Munich, Germany;

yingxiali@ibe.med.uni-muenchen.de (Y.L); ORCID 0000-0002-4501-5834

mansmann@ibe.med.uni-muenchen.de (U.M); ORCID 0000-0002-9955-8906

hornung@ibe.med.uni-muenchen.de (R.H); ORCID 0000-0002-6036-1495

Correspondence: yingxiali@ibe.med.uni-muenchen.de

Abstract: In the last decade, the possibility of using multi-omics data, that is several types of omics data available for the same patients, to predict clinical outcomes has become a popular research topic. Multi-omics data have been seen to offer the potential of outperforming single-omics data in terms of predictive performance. Nevertheless, obtaining large numbers of omics data types is complex and costly, which is why it is beneficial to only collect omics data types that contribute to improving the predictive performance. It is, however, unclear, which combinations of omics data tend to work well and which types of omics data correspondingly can, in general, be left out without hurting predictive performance. The aim of this paper was to provide answers to these questions through a large-scale benchmark study using real data. We compared the predictive performance of all 31 possible combinations of 5 types of genomic data using 14 publicly available cancer datasets with survival outcome. The considered data types were mRNA, miRNA, methylation, mutation, and copy-number variation data. Clinical data were included and prioritized in every prediction model. As prediction methods, we used random survival forests, block forests, Lasso, priority-Lasso, and IPF-LASSO. These methods were compared using repeated 5-fold cross validation. The integrated Brier score and Harrell's concordance index were used as performance metrics. To investigate the stability of the results, bootstrap analysis at the level of the included datasets was performed. Contrasting our expectations, we did not generally see an improved predictive performance by combining several omics data types. Instead, using only mRNA data or a combination of mRNA and miRNA data was sufficient in most cases. Combinations of larger numbers of omics data types tended to lead to a worsening of predictive performance. While the number of datasets included in our study is relatively large it is still limited, which is why our results must be interpreted with caution. Nevertheless, they strongly indicate that integrating many omics data types in a multi-omics prediction context can be counterproductive.

Keys: Multi-omics data, Prediction, TCGA, Benchmark, Cancer, Survival analysis

1. Introduction:

Cancer is a global public health problem due to its high morbidity and mortality [1]. It is associated with alterations in genes that control normal cell growth and death. Thus, understanding and exploiting the molecular basis of cancer has many benefits, including the possibility of building prediction models [2,3], discovering biomarkers [4], identifying abnormal pathways [5], and determining optimal treatment options.

Today, various types of omics data exist. These include genomics, epigenomics, transcriptomics, proteomics, and metabolomics data. Many of these data types are publicly available on The Cancer Genome Atlas (TCGA) [6]. In the following, the different types of molecular data are referred to as "blocks". Omics data have been used to develop predictive models for more than 20 years. These models traditionally used only one block, frequently the RNA or the mRNA block. As a well-known example, mRNA data have often been found to be useful for predicting survival response to therapy in cancer patients. With the increasing availability of other types of blocks, the focus has shifted towards constructing predictive models based on multi-omics data, that is, several block types available for the same patients. Several works have suggested that multi-omics data outperform single-omics data for generating predictive models [7–10].

For example, Li et al. [11] found that using multi-omics data delivered notably better results than using single-omics data in the prediction of lung adenocarcinoma stage.

Although the use of multi-omics data for prediction has been a well-studied topic, it has to date not been examined, which blocks consistently improve the predictive performance and which blocks can correspondingly, in general, be left out to save efforts and costs. In fact, it is quite costly and laborious to obtain different types of omics data for the same patient. The huge data volume of multi-omics data can also lead to long computation times and large consumptions of computational memory. Moreover, given the costs and efforts needed to obtain multi-omics data, when including multiple types of omics data at the same time the sample sizes associated with these data can be expected to be small. Thus, ideally, the number of blocks should be small to reduce labor and cost.

Several studies have compared the predictive performance of different block combinations; however, these studies tend to be limited in scope and yielded inconsistent results. A pioneering study by Zhao et al. [12] conducted a comparative study considering four types of genomic data. They used four datasets and prediction methods that do not take the multi-omics structure into account (more sophisticated methods were not yet available at the time of conduction of their analysis). It was observed in that study that once mRNA data and clinical covariates were included in the model, the addition of any further genomic data types did not substantially improve prediction results. Gómez-Rueda et al. [13] considered four blocks, but these were not available for the same patients, which is why these data were not multi-omics data. They performed predictive modelling using each block individually and concurrently using all blocks taken together. Here, first, the integration of four blocks produced a slightly better model performance than using any of the single blocks, and second, for the individual blocks, the best predictions were obtained using mRNA, followed by miRNA and copy-number variation (CNV).

In this paper, by aims of a large-scale benchmark experiment we aim to address whether all blocks in multi-omics data are necessary for prediction purposes or whether, in contrast, some blocks can be left out without hurting the predictive performance notably. We compared the predictive performance of all possible 31 combinations of 5 types of genomic data (mRNA, miRNA, methylation, mutation, and CNV) using 14 cancer datasets with survival outcome.

2. Methods:

2.1. Datasets

The 14 included multi-omics datasets were the same as those studied in [2], except that we included methylation data in addition. For each cancer type, there are five molecular data types and clinical data, that is, six groups of features. An overview of these 14 datasets is given in Table 1.

Table 1. Overview of the considered datasets. The second to the seventh column show the numbers of features in the respective feature blocks (clin: clinical features, cnv: CNV, mirna: miRNA, mut: mutation, met: methylation, rna: mRNA). The last four columns show, in this order, the total number of features (f), the numbers of observations (n), the numbers of uncensored observations (e), and the proportions of uncensored observations (r_e).

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

2.2. Feature selection

The permutation-based variable importance measure of random survival forests (RF-VI) can be used to rank the features with respect to their importance to prediction. It can be used in feature selection by retaining the best-ranking variables. In a previous work, we conducted a benchmark study of feature selection strategies for multi-omics data with binary outcome, where we found that RF-VI was quite robust with respect to the number of features selected and was relatively fast [14].

For blocks with more than 2500 variables, we used the RF-VI feature selection method to perform feature selection on the training dataset within cross-validation, selecting the 2500 features with the largest variable importance measure values. This was done for computational efficiency and because most variables do not carry information in the ultra-high-dimensional multi-omics data types. Because of the large numbers of features for some blocks (in particular the methylation block), we used 10000 trees per random survival forest instead of the number 500 that is default in the R package ranger (version 0.14) we used.

2.3. Prediction methods

Random survival forests (rsf) [15] are a variant of random forests [16] for survival outcomes. Random forests are ensemble classifiers that use randomly selected training samples and randomly selected subsets of variables to produce multiple, heterogeneous decision trees. They have become popular due to their ability to capture complex patterns of dependencies between the outcome and the input features. However, they are not designed to take the multi-omics group structure into account.

The block forest (bf) algorithm [3] is a variant of random forests that modifies the split point selection of random forests to incorporate the block structure of multi-omics data.

The least absolute shrinkage and selection operator (lasso) [17] is a penalized regression method that applies an L1 penalty to shrink coefficients of features without strong impact on the predictions to zero. When using multi-omics data to predict clinical outcomes, lasso regression penalizes equally each feature across all blocks by using a single penalization parameter for the entire dataset. That is, as rsf, the method does not take the multi-omics group structure into account.

The IPF-LASSO [18] is an extension of the lasso that takes the group structure into account by using different penalty parameter values for each block. We used a variant of the integrative lasso, the two-step integrative lasso with penalty factors (ipflasso) [19], which performs an efficient two-step procedure to optimize the penalty parameter values.

Priority lasso (prioritylasso) [20] is, as the ipflasso, an extension of the lasso. It is based on the principle of defining a priority order on the blocks of variables. Subsequently, prioritylasso successively fits lasso regression models to the blocks in the order of their priority, where at each step, the resulting linear predictor is used as an offset for the lasso model fit to the next block. For the current study, we, however, did not have any substantial domain knowledge needed for assigning the priority order to the blocks for the different datasets. Therefore, we used the ranking of the penalty factor values determined in the first step of the ipflasso as a surrogate for knowledge-based prioritization, that is, the block with the smallest penalty factor was given the highest priority, the block with the second smallest penalty factor the second highest priority, and so on.

2.4. Experimental settings

Clinical covariates carry important predictive information and several studies have demonstrated that their inclusion improves predictive performance [2,3]. It is important to prioritize the clinical covariates over the omics blocks to utilize their predictive information because there are typically many more omics features than clinical covariates [21]. Therefore, except for in the case of ipflasso, for which this was not possible, we prioritized the clinical covariates for all prediction methods. For rsf, this was achieved by adding all clinical variables to the mtry randomly sampled covariates for each split in the trees constituting the rsf. For bf, similarly the clinical block was always included in the blocks considered for splitting. For lasso, the coefficients of the clinical covariates were exempt from the L1 penalization-based shrinkage. Finally, for prioritylasso, the clinical block always had highest priority and, as in the case of lasso, no shrinkage was performed for the clinical covariates.

Our goal was to assess whether all omics blocks are necessary to achieve optimal predictive performance or whether there are instead specific subsets of blocks which work sufficiently well or even better than the combination of all blocks. For each dataset, we considered all $2^5 - 1 = 31$ possible combinations of the omics blocks (the clinical features were always included) and compared the predictive performances achieved with each combination. We repeated the analysis for each of the five considered prediction methods.

The integrated Brier (ibrier) score and the concordance index (cindex) were used to evaluate the predictive performance. Here, the ibrier is a calibration measure, which assesses how accurate the predicted survival functions are. In contrast, the cindex is a measure of discrimination only. It assesses, how well the prediction rule can rank different patients according to their risk. As an evaluation scheme, we used 5-fold cross-validation repeated five times. The benchmark experiment was conducted using R version 4.1.2 [22]. All R code written to produce and evaluate our results is available on GitHub (https://github.com/yingxiali/multi-omics-data, accessed on August 22, 2022).

3. Results

For reasons of clarity, we present here only the results obtained for the ibrier with rsf, bf, and ipflasso. The results obtained for the ibrier with lasso and prioritylasso and all results obtained for the cindex are shown in the supplementary material.

3.1. Ranking of the predictive information contained in all block combinations per prediction method

Figure 1 shows, separately per prediction method, for each block combination which ranking it achieved among all 31 possible block combinations per dataset. Supplementary material Figures S5 and S6 show, separately per prediction method, the distribution of the mean cross-validated ibrier and cindex values across the datasets for all 31 possible block combinations.

The results differ quite considerably across the different prediction methods. However, a consistent observation we can make is that the best performances were achieved with only few blocks. Adding more blocks did not deliver better predictive performance, but in contrast, tended to deliver worse results. For rsf and bf, we see that mRNA was very important to prediction as the block combinations that performed best all included mRNA. Apart from the latter specific observation, there is no clear picture with respect to the importance of each individual block. In general, the boxplots in Figure 1 reveal that the results differ quite strongly across the datasets, in particular in the case of ipflasso. The results obtained for lasso and prioritylasso are shown in Supplementary Figure S1. Interestingly, lasso was the only method for which using more blocks tended to deliver better prediction results. For prioritylasso, we again see a clear trend towards worse predictive performance for block combinations with many blocks, where the best results were obtained with single blocks. In the next subsection it will, however, be seen that using prioritylasso tended to lead to worse prediction results than the other prediction methods. While we do see differences in the results obtained for the cindex (Supplementary Figures S2 and S3), the general conclusions are very similar to those obtained with the ibrier. Exceptions are that for lasso we do no longer observe a trend towards better predictive importance by including more blocks and that for ipflasso there was less variability of the results across datasets.

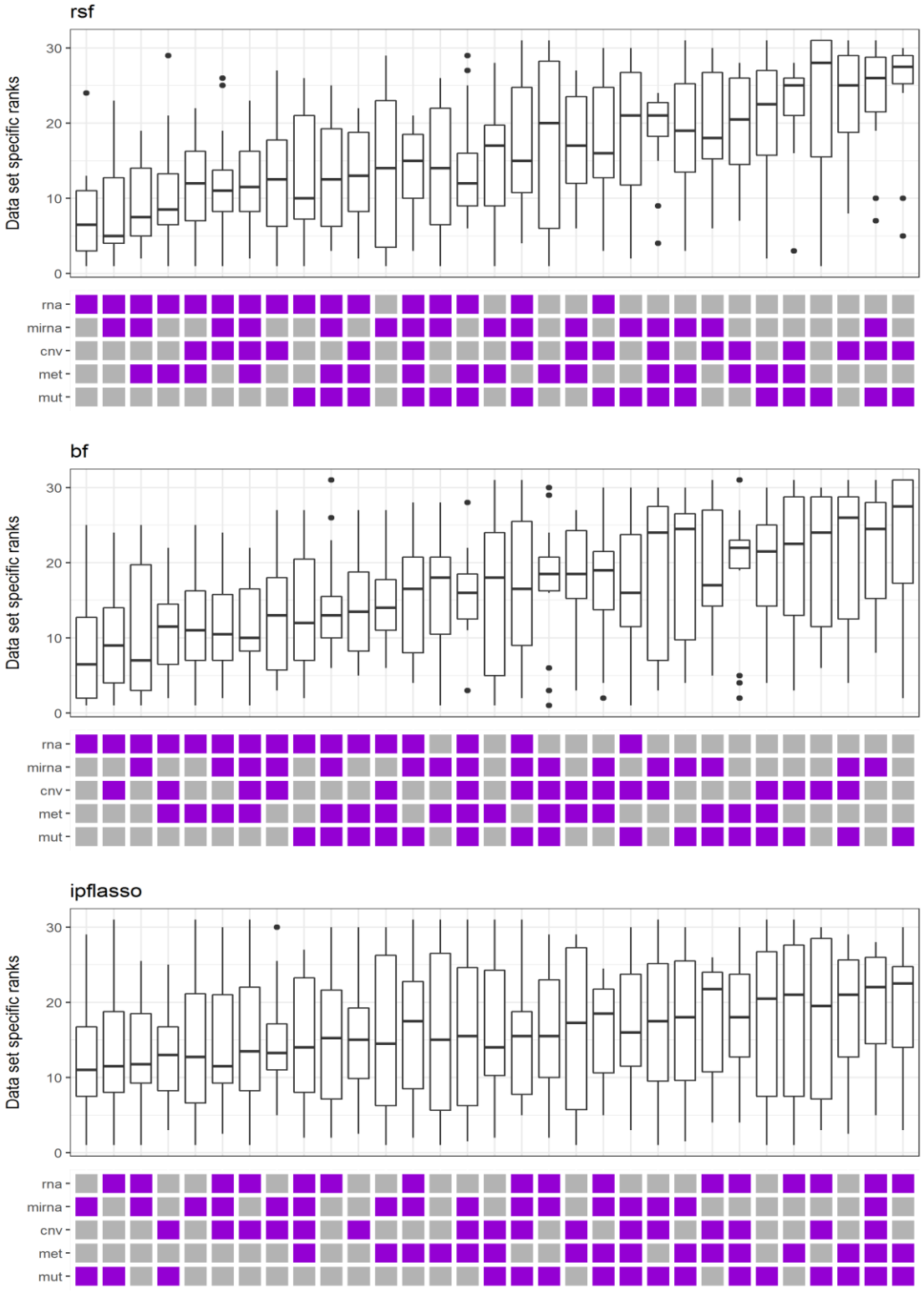

Figure 1 Dataset specific ranks of each block combination among all block combinations in terms of the cross-validated ibrier values: rsf, bf, and ipflasso. Smaller ranks indicate a better predictive performance. The block combinations are sorted in increasing order according to the mean ranks across the datasets. cnv: CNV, mirna: miRNA, mut: mutation, met: methylation, rna: mRNA.

Even though the number of datasets included in our benchmark experiment is comparably large, we still must consider that the mean ranks obtained for the block combinations are associated with considerable variability. The large variances observed in the boxplots showing the results obtained for the different datasets was already indicative of this. To investigate this variability, we performed bootstrap analysis [23,24] at the level of the 14 included datasets. For each sub-analysis (e.g, those shown in Figure 1 and Supplementary Figure S1), we drew 5000 bootstrap samples, and each time re-calculated the mean ranks and the ranks of each mean rank among all other mean ranks. These ranks of the mean ranks will be denoted 'positions' in the following. Subsequently we calculated the mean positions across the 5000 bootstrap samples and 95% percentile confidence intervals for these mean positions. These intervals are calculated by taking the 2.5% quantile and 97.5% quantile of the 5000 positions calculated using the bootstrap samples.

The results corresponding to Figure 1 are shown in Table 2. As expected from the high variability observed across the datasets, the confidence intervals are wide in all cases. However, for rsf and bf our general conclusions made above can be confirmed. It is seen that the upper bounds of the confidence intervals obtained for the best-performing combinations that tended to include only few blocks are still relatively low in most cases. This confirms that these are indeed among the better combinations. For example, the confidence intervals obtained for using only mRNA suggest that these are among the ten best combinations. For ipflasso, however, the confidence intervals are very wide, which is why for these we cannot confirm that the best combinations are significantly better. However, for all three methods (rsf, bf, and ipflasso) the lower bounds of the confidence intervals obtained for the combination of all five blocks is larger than five, which confirms that using all blocks does in general not lead to the best predictive performance. The results of the bootstrap analysis for lasso and prioritylasso are shown in Supplementary Table S1. For lasso the confidence intervals tend to be very wide, which is why we cannot draw reliable conclusions on the ordering of the block combinations from this analysis. This result was expected given the large variability of the results across datasets for lasso (cf. Supplementary Figure S1). For prioritylasso the confidence intervals tended to much narrower. However, as stated above in the following subsection it will be seen that prioritylasso tended to lead to worse prediction results than the other prediction methods, which is why we do not interpret these results any further.

The results of the bootstrap analysis obtained for the cindex (Supplementary Tables S2 and S3) are quite similar to those obtained for the ibrier. The confidence intervals tend to be slightly narrower. For all prediction methods with exception of ipflasso, in the case of the combination of all blocks, the values of the mean positions and the lower bounds of their confidence intervals are much higher for the cindex. This could indicate that the cindex suffers more from including unnecessary blocks. Note that we did not adjust the confidence intervals for multiple testing. This choice was made because, in the interpretation of the results, we did not consider each individual block combination. Instead, we focused on verifying few specific observations, namely whether the combinations with the best mean positions are indeed among the best positions and whether the combinations of all blocks are significantly worse than the combinations with the best positions.

Table 2 Results of the bootstrap analysis for the ibrier: rsf, bf, and ipflasso. The rows are ordered according to the positions obtained for rsf calculated using all datasets (without bootstrap). The columns "mean" and "ci" show the mean positions calculated using the 5000 bootstrap samples and their 95% percentile confidence intervals. cnv: CNV, mirna: miRNA, mut: mutation, met: methylation, rna: mRNA.

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]

3.2. Ranking of the predictive performance of all prediction methods on all block combinations

In the previous subsection we analyzed the results per prediction method. This analysis did not yet allow to judge which combinations of prediction methods and blocks tend to deliver the best prediction results. Figure 2 shows, per dataset, the ranking each combination of prediction method and blocks achieved among all 155 combinations of prediction method and blocks. For reasons of clarity, we only show the 30 combinations with the smallest positions. The corresponding results for cindex are shown in Supplementary Figure S4.

The prediction method bf occurred the most often in the 30 best combinations and rsf, lasso, and ipflasso occurred roughly the same numbers of times in these combinations.

The method prioritylasso was not featured in the best combinations. Almost all best combinations featured mRNA and the two best combinations only used mRNA. In contrast, mutation was featured only infrequently among the best combinations. Nevertheless, there is again a large variability between the results obtained for the different datasets. Interestingly, for the cindex (Supplementary Figure S4) lasso was featured by far the most frequently in the 30 best combinations. This result seems surprising at first given that lasso was among the worst-performing methods in the benchmark studies of [2] and [3]. However, in contrast to these preceding benchmark studies we did not penalize the coefficients of the clinical covariates. This likely explains, why in our benchmark study lasso performed much better given the high predictive importance of clinical covariates. A disadvantage of lasso seen in Figure 2 and Supplementary Figure S4 is that it tends to require more blocks than the other methods. The majority of the 30 best combinations featured mRNA also for the cindex. It is again important to not over-interpret details of the obtained results because the variability across the different datasets is large here as well.

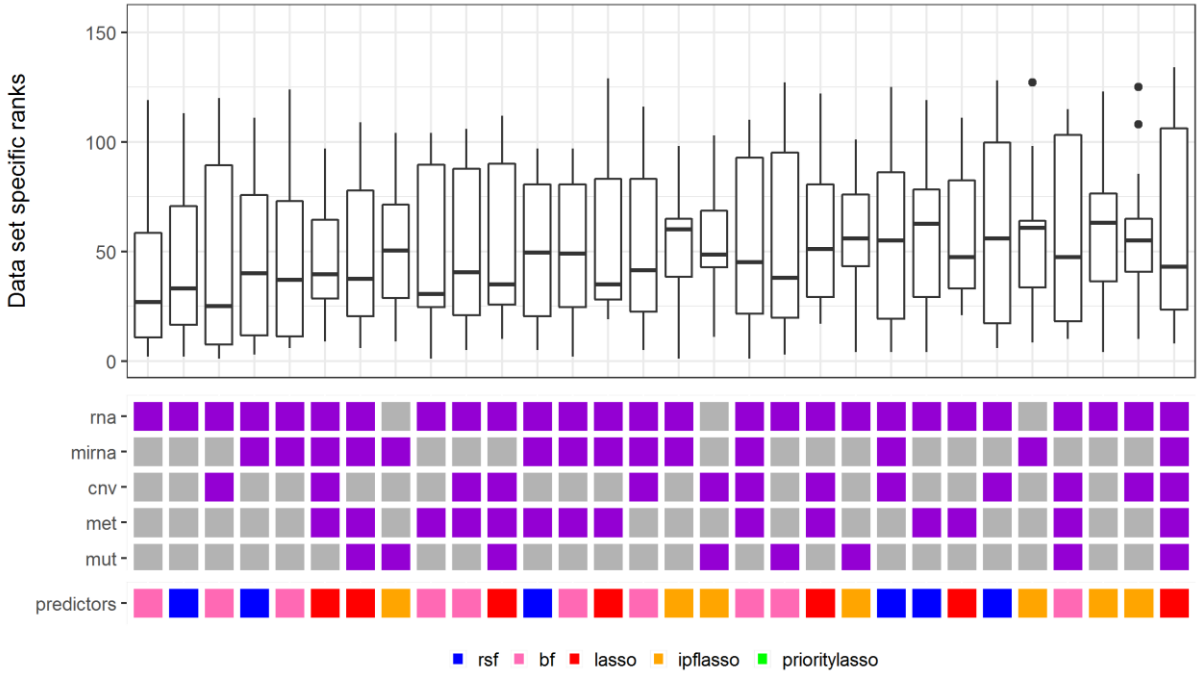

Figure 2 Dataset specific ranks of each combination of prediction method and blocks among all 155 combinations of prediction method and blocks in terms of the cross-validated ibrier values. Smaller ranks indicate a better predictive performance. The combinations are sorted in increasing order according to the mean ranks across the datasets. For reasons of clarity only the 30 combinations with the smallest positions are shown. cnv: CNV, mirna: miRNA, mut: mutation, met: methylation, rna: mRNA.

3.3. The best-performing combinations of prediction methods and blocks per dataset

As seen above, the ranks the different combinations of prediction methods and blocks achieved varied strongly between the datasets. It is interesting to learn about which prediction methods and block combinations are most successful for which datasets. Table 2 shows, for each dataset, the combinations of prediction method and blocks associated with the smallest cross-validated ibrier values and largest cross-validated cindex values. For the great majority of datasets, the best performance was achieved using only up to

two blocks. We observed quite large variability of the block combinations used across the datasets and also for each dataset, between the two performance measures. While it is of course not clear how much of this is due to random variation, it is congruent with the observation made in the previous subsections that there is great variability of the ranks of the block combinations across datasets. For more than half of the datasets, mRNA was used and miRNA took the second place. For the cindex, fewer blocks tended to be used than for the ibrier. This could indicate that for achieving good calibration more blocks tend to be needed than for achieving good discrimination only. In the case of the cindex, for five datasets only mRNA was used. Another difference observed between ibrier and cindex is that methylation data was used quite frequently in the case of the ibrier, but only for one dataset in the case of the cindex. Regarding the used prediction methods, we do not see a clear winner. For both performance measures, each prediction method was used at least for one dataset.

Table 3. The combinations of prediction methods and blocks associated with the smallest cross-validated ibrier values and largest cross-validated cindex values – separately for each dataset. cnv: CNV, mirna: miRNA, mut: mutation, met: methylation, rna: mRNA.

	ibrier		**cindex**	
dataset	**prediction method**	**blocks**	**prediction method**	**blocks**
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

4. Discussion:

Even though we saw strong variability of the results across datasets, prediction methods, and performance metrics, we were able to make important general observations. The prediction rules obtained based on the combination of all available omics blocks consistently performed worse than combinations with few blocks. Thus, contrary to current practice in prediction using multi-omics data it is not advisable to use as many omics data types as possible. Instead, in most cases prediction rules based on single or two omics data types seem to perform best.

In our analysis, mRNA was included in the best combinations the most frequently, followed by miRNA. This observation can be well interpreted biologically. Measurements of mRNA and miRNA at the transcription level affect clinical outcomes in cancer more

directly than molecular features (CNV, mutation and methylation) measured at the DNA/epigenetic level. In contrast, other genomic measurements affect clinical outcomes by influencing gene expression (mRNA and miRNA). Therefore, measurements of transcript levels may carry the richest information on prognosis and mutation, methylation, and CNV data may not provide much additional predictive power. Published studies have illustrated that multi-omics data are important for understanding cancer biology [25,26]. However, our large-scale benchmark study strongly suggests that, for prediction, integrating many omics data types can hurt predictive performance. Apart from the reasons given above, another factor contributing to this worsening of the predictive performance could be that large models based on many variables tend to be less stable.

Herrmann et al. [2] found that the predictive performance of multi-omics data using state-of-the-art prediction models is limited in their benchmark study. Here, only one prediction method based on multi-omics data (slightly) outperformed the clinical model. However, Herrmann et al. [2] always used all blocks for prediction, which, given the result of the current paper, could in part explain why multi-omics prediction performed so poorly in their study.

By prioritizing the clinical covariates, we exploited the predictive information contained in them well. Given that the predictive information contained in the clinical covariates and the omics features is overlapping it might be assumed that, if we had not prioritized the clinical covariates, more omics data types would have been necessary to achieve optimal predictive performance. However, this seems unlikely because few blocks were necessary for almost all datasets and we made the same observation in the case of ipflasso, the only method for which we did not prioritize the clinical covariates. Irrespective of this, it is always important to prioritize the clinical covariates to exploit their strong predictive information.

As seen in Figures 1 and 2, the ranks of the different block combinations varied strongly between different datasets. Consequently, as seen in Table 3, for different datasets different block combinations were optimal. For some datasets, the optimal block combinations did neither include mRNA nor miRNA. Thus, no one combination of blocks is better than all other combinations for all datasets. This large variability emphasizes the importance of large-scale benchmark studies using many datasets, as performed in this paper. It is well known that many observations are necessary to draw valid statistical conclusions, which is due to the large variability between these observations. However, this issue is often overlooked when designing benchmark experiments where the datasets play the roles of the observations [27]. It is common in published benchmark studies that only few (e.g, 5 to 7) datasets are considered.

The ranks of the different block combinations also varied quite strongly between the considered prediction methods. Here, we did, however, not observe structural differences between methods that do (bf, ipflasso, prioritylasso) and do not consider the group structure of the multi-omics data (rsf, lasso). The best-performing prediction rules (Figure 2) also included many which were obtained by prediction methods that do not consider the group structure of the multi-omics data. In contrast, in the large-scale benchmark studies by Herrmann et al. [2] and Hornung et al. [3] most prediction methods that consider the group structure outperformed those that do not. This discrepancy

can likely be explained by the fact that we prioritized the clinical covariates also for those methods that do not consider the group structure (cf. Section 2.4), which was not done in Herrmann et al. [2] and Hornung et al. [3]. In addition, Nießl et al. [28] have shown that the results of benchmark studies are in general variable and are sensitive to analytic choices even if large numbers of datasets are used.

Lastly, the results also vary between the two considered performance metrics. The cindex is not a strictly appropriate scoring rule [29], as it only measures discrimination. In contrast, the ibrier measures how well the predictions match the true outcomes values. Therefore, the ibrier should be considered as the primary measure of predictive accuracy.

Given the strong variability across datasets it is difficult to judge, how strongly the aggregated results are affected by random variation. We took great care not to interpret details of the obtained results but focused on general observations that could be made across the different prediction methods and performance metrics. Moreover, using bootstrap analysis we were able to strengthen our main conclusions.

5. Conclusions:

The use of multi-omics data to predict clinical outcomes has been an active and productive area of research in recent years. For understanding cancer biology, it is important to combine several different omics data types to multi-omics data. However, obtaining such data is complex and costly, which is why for prediction purposes it would be beneficial to only collect omics data types that contribute to improving the predictive performance.

In the large-scale benchmark study presented in this paper we found that combining many different types of omics data can hurt the performance of multi-omics prediction. Instead, our results indicate that using only few data types tends to be associated with better performance. Here, in most cases mRNA or combinations of mRNA and miRNA are sufficient, but for some datasets also other omics data types are important. As found previously [2] irrespective of the included omics data types it is important to also include clinical covariates and to prioritize them in the prediction.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

All R code written to produce and evaluate our results is available on GitHub (https://github.com/yingxiali/multi-omics-data, accessed on August 22, 2022).

Competing interests

The authors declare that they have no conflict of interest.

Funding

Y.L. was supported by the China Scholarship Council (CSC, No. 201809505004). R.H. was supported by the German Science Foundation (DFG-Einzelförderung HO6422/1-2).

Authors' contributions

Supervision, R.H. and U.M.; experimental design, Y.L. and R H.; data analysis, Y.L.; writing—original draft preparation, Y.L.; writing—review and editing, R.H. and U.M.; All authors have read and agreed to the published version of the manuscript.

References:

1. Ferlay, J.; Soerjomataram, I.; Dikshit, R.; Eser, S.; Mathers, C.; Rebelo, M.; Parkin, D.M.; Forman, D.; Bray, F. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int. J. cancer* **2015**, *136*, E359--E386.
2. Herrmann, M.; Probst, P.; Hornung, R.; Jurinovic, V.; Boulesteix, A.-L. Large-scale benchmark study of survival prediction methods using multi-omics data. *Brief. Bioinform.* **2021**, *22*, bbaa167.
3. Hornung, R.; Wright, M.N. Block Forests: random forests for blocks of clinical and omics covariate data. *BMC Bioinformatics* **2019**, *20*, 1–17.
4. Mariani, M.; He, S.; McHugh, M.; Andreoli, M.; Pandya, D.; Sieber, S.; Wu, Z.; Fiedler, P.; Shahabi, S.; Ferlini, C. Integrated multidimensional analysis is required for accurate prognostic biomarkers in colorectal cancer. *PLoS One* **2014**, *9*, e101065.
5. Chari, R.; Coe, B.P.; Vucic, E.A.; Lockwood, W.W.; Lam, W.L. An integrative multi-dimensional genetic and epigenetic strategy to identify aberrant genes and pathways in cancer. *BMC Syst. Biol.* **2010**, *4*, 1–14.
6. Tomczak, K.; Czerwińska, P.; Wiznerowicz, M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol.* **2015**, *19*, A68.
7. Yong, Z.; Dun-wei, G.; Wan-qiu, Z. Feature selection of unreliable data using an improved multi-objective PSO algorithm. *Neurocomputing* **2016**, *171*, 1281–1290, doi:10.1016/j.neucom.2015.07.057.
8. Dong, Y.; Yang, W.; Wang, J.; Zhao, J.; Qiang, Y.; Zhao, Z.; Kazihise, N.G.F.; Cui, Y.; Yang, X.; Liu, S. MLW-gcForest: A multi-weighted gcForest model towards the staging of lung adenocarcinoma based on multi-modal genetic data. *BMC Bioinformatics* **2019**, *20*, 1–14, doi:10.1186/s12859-019-3172-z.
9. Sun, D.; Li, A.; Tang, B.; Wang, M. Integrating genomic data and pathological images to effectively predict breast cancer clinical outcome. *Comput. Methods Programs Biomed.* **2018**, *161*, 45–53, doi:10.1016/j.cmpb.2018.04.008.
10. Kim, D.; Shin, H.; Song, Y.S.; Kim, J.H. Synergistic effect of different levels of genomic data for cancer clinical outcome prediction. *J. Biomed. Inform.* **2012**, *45*, 1191–1198, doi:10.1016/j.jbi.2012.07.008.
11. Li, Y.; Mansmann, U.; Du, S.; Hornung, R. Synergistic Effects of Different Levels of Genomic Data for the Staging of Lung Adenocarcinoma: An Illustrative

Study. *Genes (Basel).* **2021**, *12*, 1872.

12. Zhao, Q.; Shi, X.; Xie, Y.; Huang, J.; BenShia, C.; Ma, S. Combining multidimensional genomic measurements for predicting cancer prognosis: Observations from TCGA. *Brief. Bioinform.* **2015**, *16*, 291–303, doi:10.1093/bib/bbu003.

13. Gómez-Rueda, H.; Martínez-Ledesma, E.; Martínez-Torteya, A.; Palacios-Corona, R.; Trevino, V. Integration and comparison of different genomic data for outcome prediction in cancer. *BioData Min.* **2015**, *8*, 1–12, doi:10.1186/s13040-015-0065-1.

14. Li, Y.; Mansmann, U.; Du, S.; Hornung, R. Benchmark study of feature selection strategies for multi‐omics data. *BMC Bioinformatics* **2022**, 1–18, doi:10.1186/s12859-022-04962-x.

15. Ishwaran, H.; Kogalur, U.B.; Blackstone, E.H.; Lauer, M.S. Random survival forests. *Ann. Appl. Stat.* **2008**, *2*, 841–860, doi:10.1214/08-AOAS169.

16. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32.

17. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* **1996**, *58*, 267–288.

18. Boulesteix, A.L.; De Bin, R.; Jiang, X.; Fuchs, M. IPF-LASSO: Integrative L1-Penalized Regression with Penalty Factors for Prediction Based on Multi-Omics Data. *Comput. Math. Methods Med.* **2017**, *2017*, doi:10.1155/2017/7691937.

19. Schulze, G. Clinical outcome prediction based on multi-omics data: extension of IPF-LASSO, MA thesis. Munich: Ludwig-Maximilians-University. Department of Statistics, 2017.

20. Klau, S.; Jurinovic, V.; Hornung, R.; Herold, T.; Boulesteix, A.L. Priority-Lasso: A simple hierarchical approach to the prediction of clinical outcome using multi-omics data. *BMC Bioinformatics* **2018**, *19*, 1–14, doi:10.1186/s12859-018-2344-6.

21. Boulesteix, A.-L.; Sauerbrei, W. Added predictive value of high-throughput molecular data to clinical data and its validation. *Brief. Bioinform.* **2011**, *12*, 215–229.

22. Team, R.C. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. *http//www. R-project. org/* **2013**.

23. Efron, B. Bootstrap Methods: Another Look at the Jackknife. *Ann. Stat.* **1979**, *7*, 1–26.

24. Stine, R. An introduction to bootstrap methods: Examples and ideas. *Sociol. Methods Res.* **1989**, *18*, 243–291.

25. Huang, S.C.; Clarke, D.C.; Gosline, S.J.C.; Labadorf, A.; Chouinard, C.R.; Gordon, W.; Lauffenburger, D.A.; Fraenkel, E. Linking proteomic and transcriptional data through the interactome and epigenome reveals a map of oncogene-induced signaling. *PLoS Comput. Biol.* **2013**, *9*, e1002887.

26. Heo, Y.J.; Hwa, C.; Lee, G.-H.; Park, J.-M.; An, J.-Y. Integrative Multi-Omics Approaches in Cancer Research: From Biological Networks to Clinical Subtypes. *Mol. Cells* **2021**, *44*, 433.

27. Boulesteix, A.-L.; Hable, R.; Lauer, S.; Eugster, M.J.A. A statistical framework for hypothesis testing in real data comparison studies. *Am. Stat.* **2015**, *69*, 201–212.

28. Nießl, C.; Herrmann, M.; Wiedemann, C.; Casalicchio, G.; Boulesteix, A.-L. Over-optimism in benchmark studies and the multiplicity of design and analysis options when interpreting their results. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2022**, *12*, e1441.

29. Blanche, P.; Kattan, M.W.; Gerds, T.A. The c-index is not proper for the evaluation of-year predicted risks. *Biostatistics* **2019**, *20*, 347–357.

The supplementary material of contribution 3

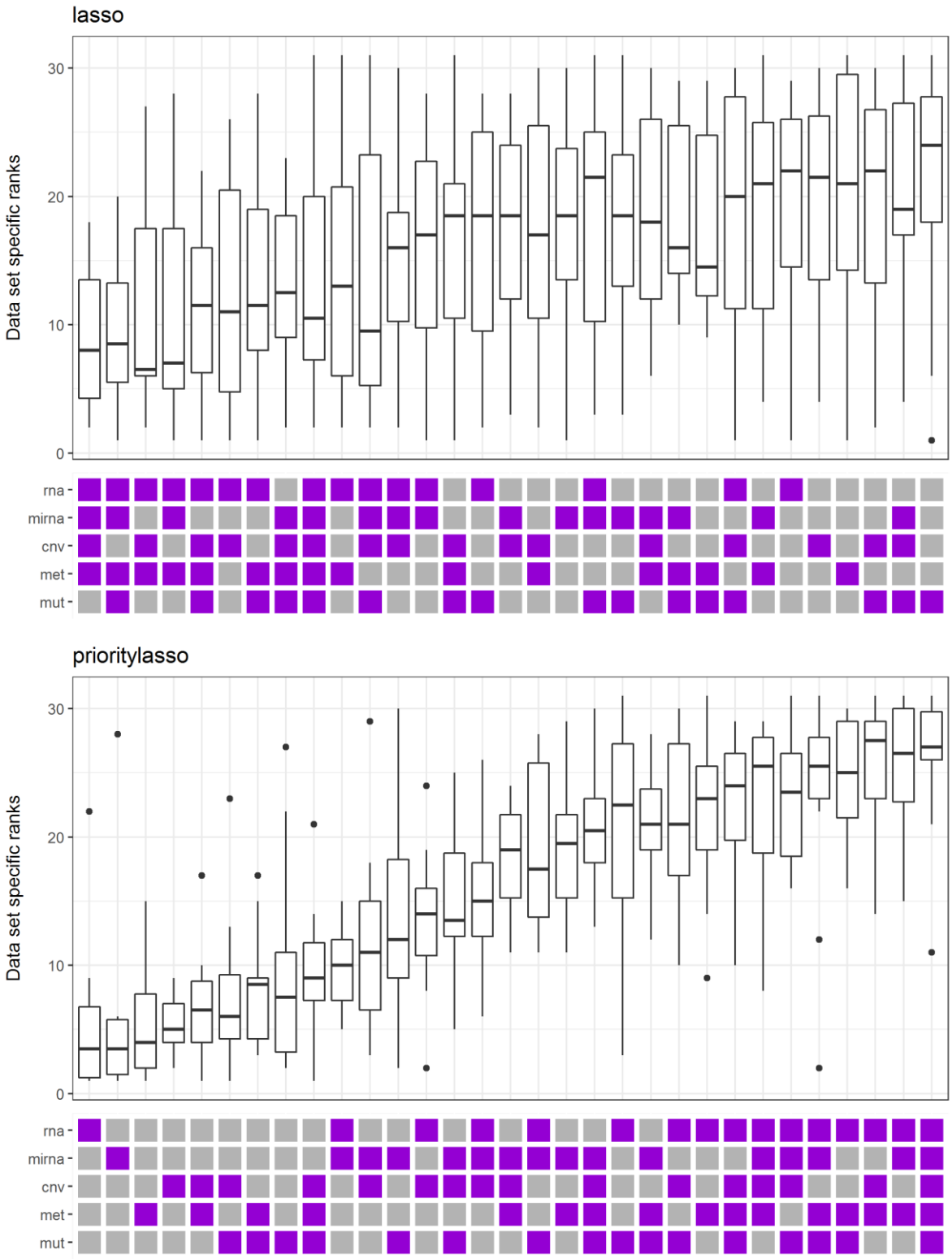

Figure S1 Dataset specific ranks of each block combination among all block combinations in terms of the cross-validated ibrier values: lasso, prioritylasso. Smaller ranks indicate a better predictive performance. The block combinations are sorted in increasing order according to the mean ranks across the datasets. cnv: CNV, mirna: miRNA, mut: mutation, met: methylation, rna: mRNA.

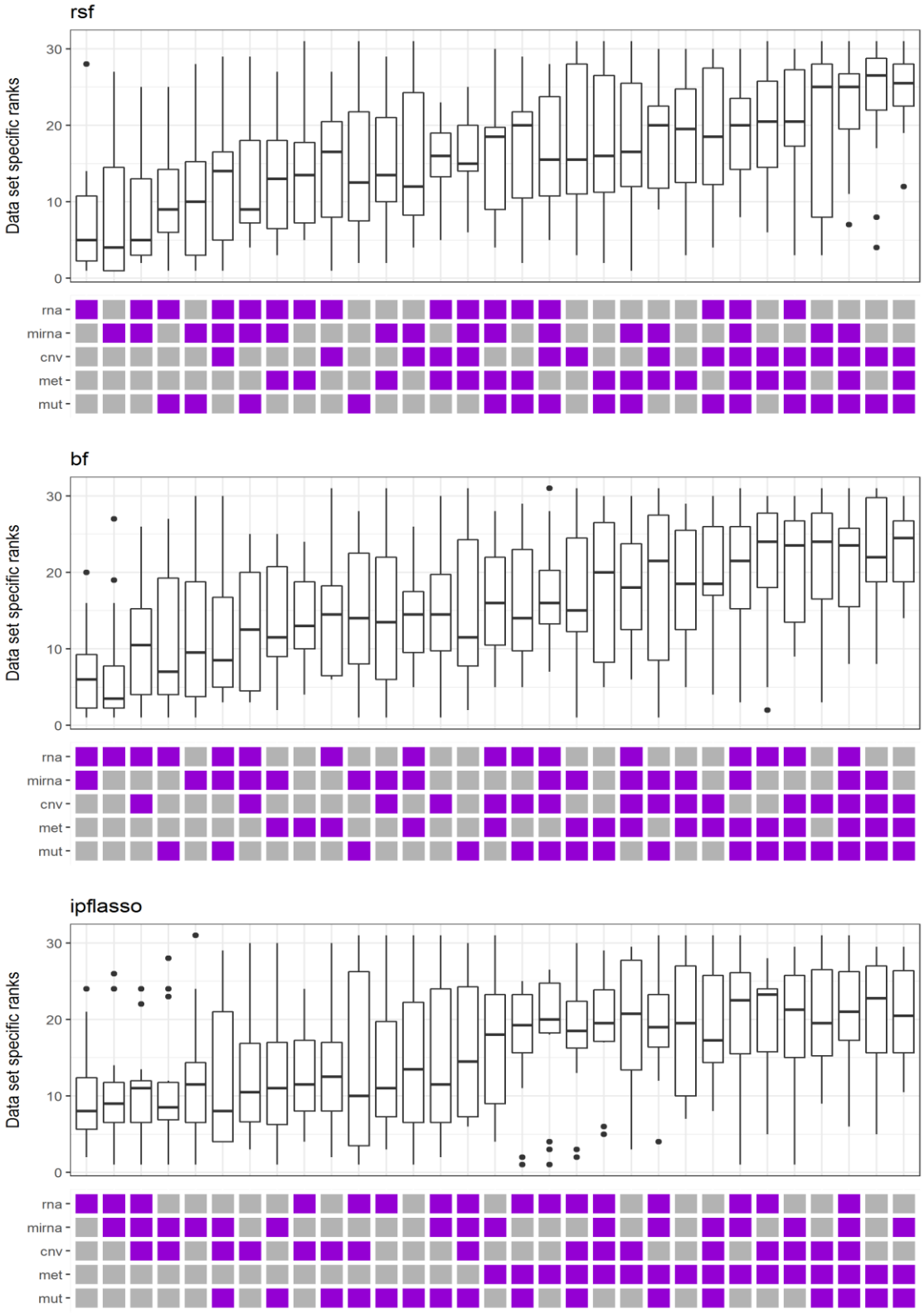

Figure S2 Dataset specific ranks of each block combination among all block combinations in terms of the cross-validated cindex values: rsf, bf, and ipflasso. Smaller ranks indicate a better predictive performance. The block combinations are sorted in increasing order according to the mean ranks across the datasets.cnv: CNV, mirna: miRNA, mut: mutation, met: methylation, rna: mRNA.

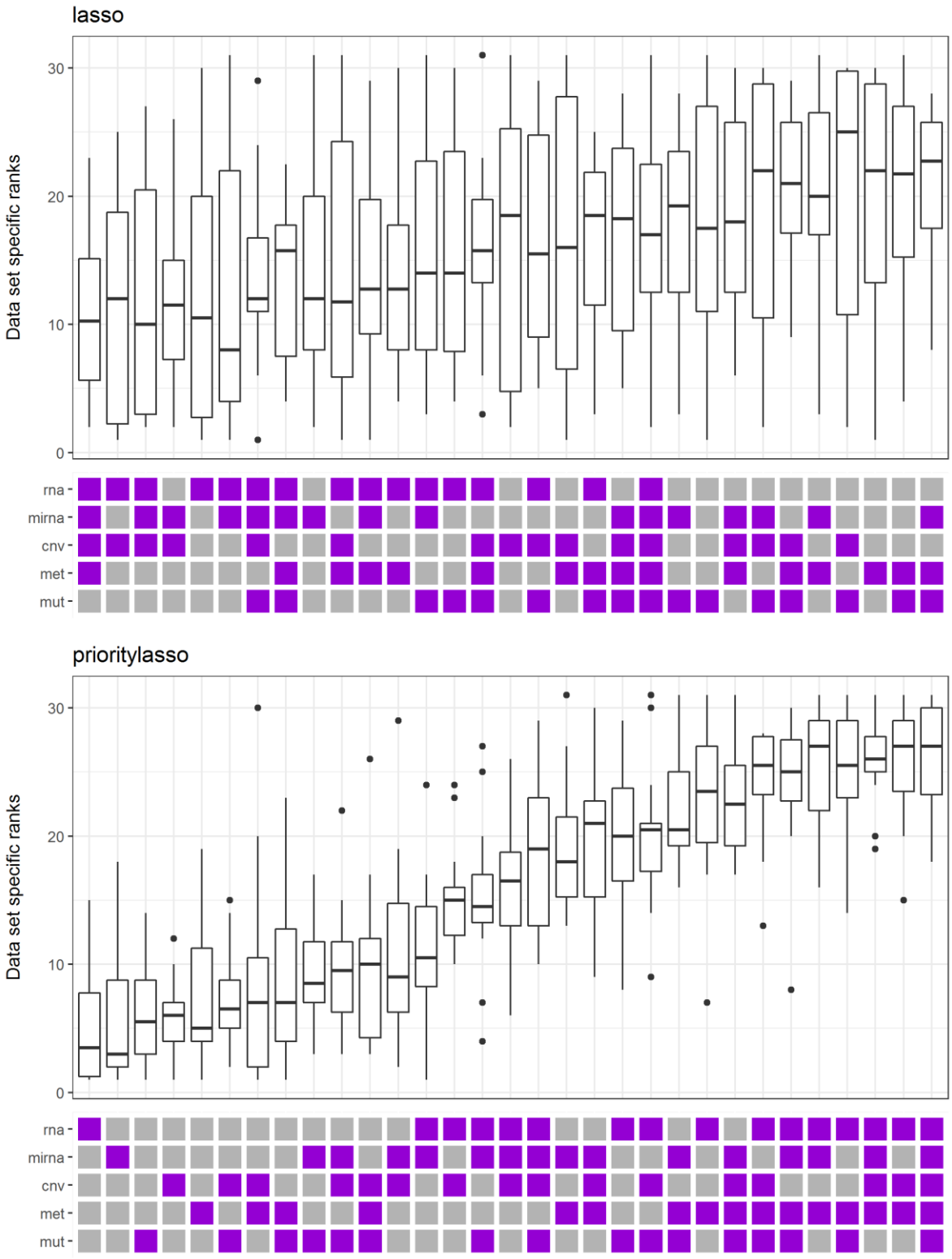

Figure S3 Data-set specific ranks of each block combination among all block combinations in terms of the cross-validated cindex values: lasso, prioritylasso. Smaller ranks indicate a better predictive performance. The block combinations are sorted in increasing order according to the mean ranks across the datasets. cnv: CNV, mirna: miRNA, mut: mutation, met: methylation, rna: mRNA.

Table S1 Results of the bootstrap analysis for the ibrier: lasso, prioritylasso. The rows are ordered according to the positions obtained for rf calculated using all datasets (without bootstrap). The columns "mean" and "ci" show the mean positions calculated using the 5000 bootstrap samples and their 95% percentile confidence intervals. cnv: CNV, mirna: miRNA, mut: mutation, met: methylation, rna: mRNA.

N0.	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]

Table S2 Results of the bootstrap analysis for the cindex: rsf, bf, and ipflasso. The rows are ordered according to the positions obtained for rsf calculated using the data (without bootstrap). The columns "mean" and "ci" show the mean positions calculated using the 5000 bootstrap samples and their 95% percentile confidence intervals. cnv: CNV, mirna: miRNA, mut: mutation, met: methylation, rna: mRNA.

N0.	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]

Table S3 Results of the bootstrap analysis for the cindex: lasso, prioritylasso. The rows are ordered according to the positions obtained for rsf calculated using all datasets (without bootstrap). The columns "mean" and "ci" show the mean positions calculated using the 5000 bootstrap samples and their 95% percentile confidence intervals. cnv: CNV, mirna: miRNA, mut: mutation, met: methylation, rna: mRNA.

N0.	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,]

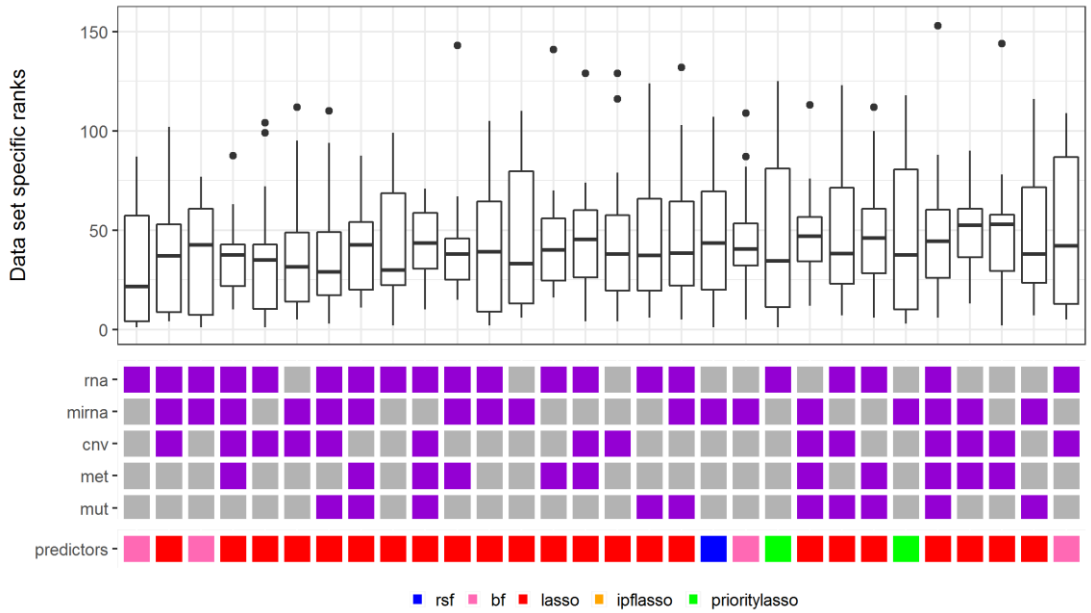

Figure S4 Dataset specific ranks of each combination of prediction method and blocks among all 155 combinations of prediction method and blocks in terms of the cross-validated cindex values. Smaller ranks indicate a better predictive performance. The combinations are sorted in increasing order according to the mean ranks across the datasets. For reasons of clarity only the 30 combinations with the smallest positions are shown. cnv: CNV, mirna: miRNA, mut: mutation, met: methylation, rna: mRNA.

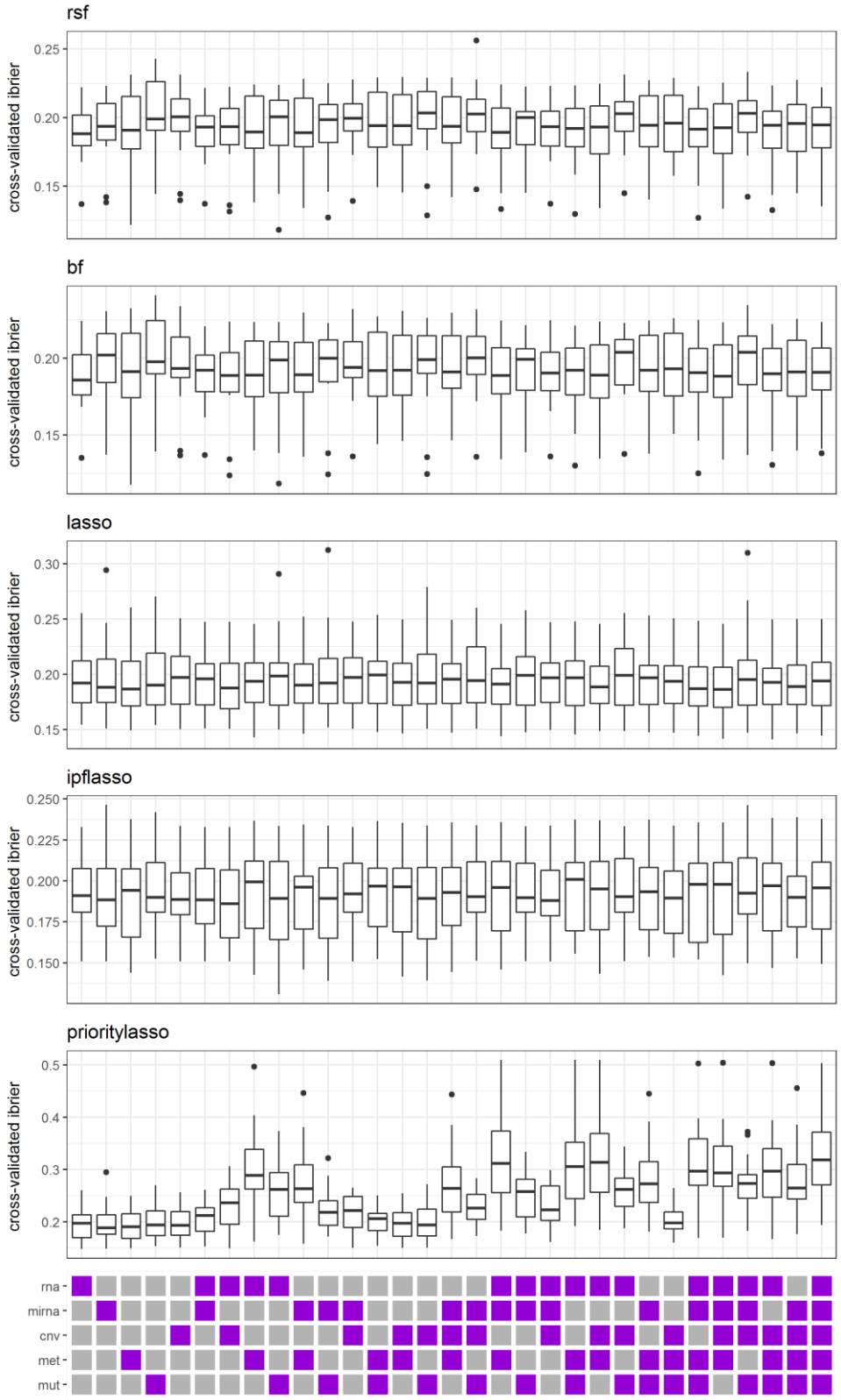

Figure S5 the distribution of the mean cross-validated ibrier values for each prediction method across the datasets for all 31 possible block combinations, respectively. cnv: CNV, mirna: miRNA, mut: mutation, met: methylation, rna: mRNA.

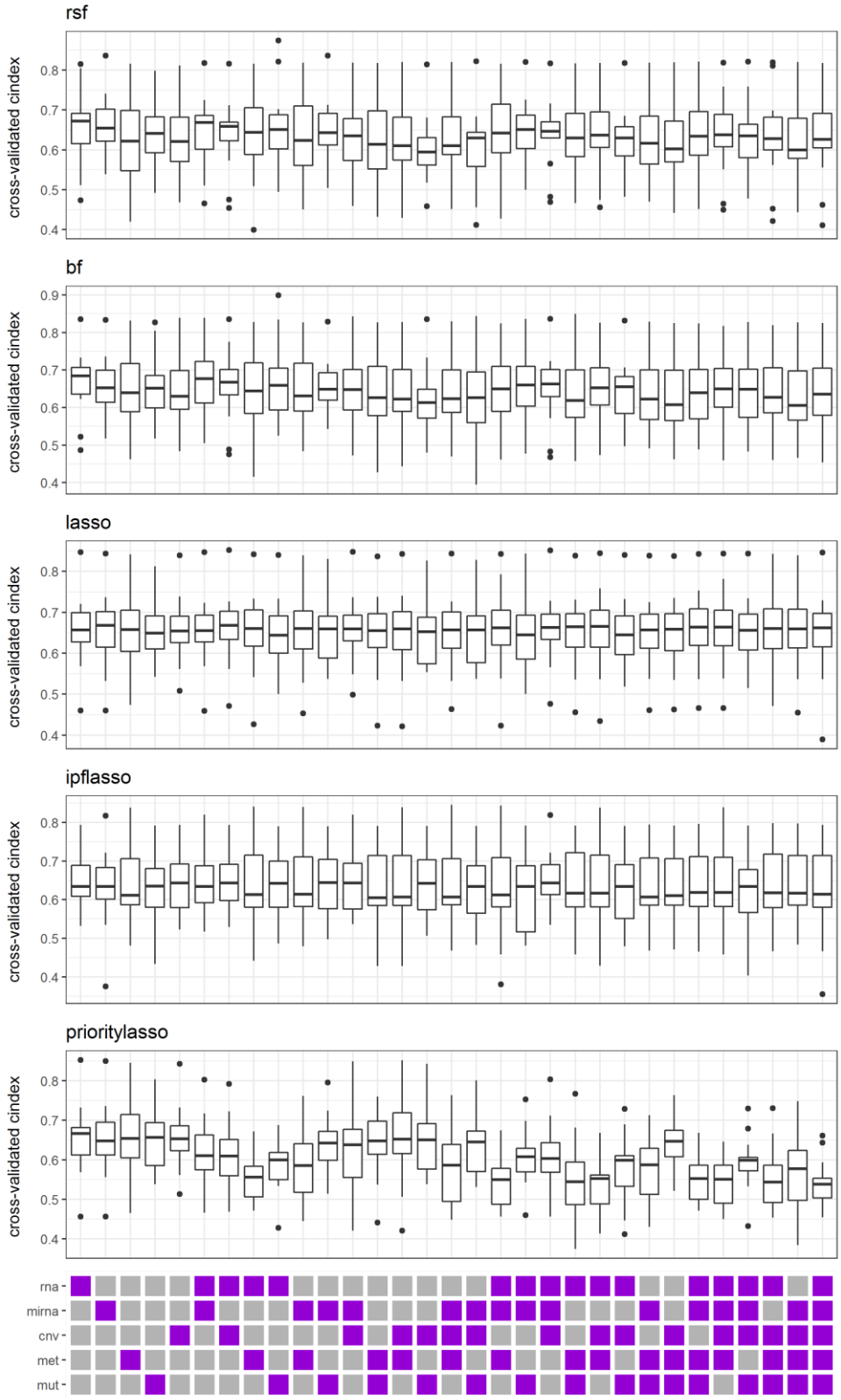

Figure S6 the distribution of the mean cross-validated cindex values for each prediction method across the datasets for all 31 possible block combinations, respectively. cnv: CNV, mirna: miRNA, mut: mutation, met: methylation, rna: mRNA.

Acknowledgements

First of all, I would like to express my deepest gratitude to Prof. Dr. Ulrich Mansmann. Thank you for giving me the opportunity to write this thesis, for always supporting me and my research with a positive attitude and great patience, and for always being available for questions and help.

Moreover, I am utmost grateful to Dr. Roman Hornung for his efforts in selecting, writing, and revising each of my papers, the successful completion of which was inseparable from his guidance and direction. Thank you for your help in each discussion and for sharing your knowledge with me.

I would like to thank my former colleagues, Jian Li, Mengying Zhang, and Shangming Du, who helped me adapt to the working environment.

I would also like to thank the China Scholarship Council (CSC) for providing me with research funding so that I can focus on my research without financial distress during my studies.

Thanks to my many Chinese friends in Germany for taking me to explore the food and beauty of Munich. Special thanks to my boyfriend for his encouragement and support during my master and PhD studies.

Finally, I sincerely thank my family for allowing me to do whatever I want to do.