bioRxiv preprint doi: https://doi.org/10.1101/2022.04.14.488153; this version posted April 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Systems Biology

Subject Section

Multi-Omics Regulatory Network Inference in the Presence of Missing Data

Juan D. Henao[1], Michael Lauber[2], Manuel Azevedo[1], Anastasiia Grekova[1], Markus List[2], Christoph Ogris[1,#], and Benjamin Schubert[1,3,#,*]

[1]Helmholtz Zentrum München, Computational Health Department, Ingolstädter Landstraße 1, 85764 Munich, Germany, Member of the German Center for Lung Research (DZL); [2]Chair of Experimental Bioinformatics, TUM School of Life Sciences, Technical University of Munich, Maximus-von-Imhof-Forum 3, 85354 Freising; [3]Department of Mathematics, Technical University of Munich, 85748 Garching bei München, Germany

* To whom correspondence should be addressed.
Joint last authors

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: A key problem in systems biology is the discovery of regulatory mechanisms that drive phenotypic behavior of complex biological systems in the form of multi-level networks. Modern multi-omics profiling techniques probe these fundamental regulatory networks but are often hampered by experimental restrictions leading to missing data or partially measured omics types for subsets of individuals due to cost restrictions. In such scenarios, in which missing data is present, classical computational approaches to infer regulatory networks are limited. In recent years, approaches have been proposed to infer sparse regression models in the presence of missing information. Nevertheless, these methods have not been adopted for regulatory network inference yet.
Results: In this study, we integrated regression-based methods that can handle missingness into KiMONo, a **K**nowledge gu**I**ded **M**ulti-**O**mics **N**etw**o**rk inference approach, and benchmark their performance on commonly encountered missing data scenarios in single- and multi-omics studies. Overall, two-step approaches that explicitly handle missingness performed best for a wide range of random- and block-missingness and noise levels, while methods implicitly handling missingness performed worst and were generally unstable. Our results show that robust multi-omics network inference with KiMONo is feasible and thus allows users to leverage available multi-omics data to its full extent.
Availability: https://github.com/cellmapslab/kimono
Contact: benjamin.schubert@helmholtz-muenchen.de
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Complex biological systems are organized in multi-level, dynamically controlled networks that regulate and maintain the phenotypic behavior of individual cells and their response to environmental changes (Romero *et al.*, 2012). Uncovering these multi-level networks and systemically understanding the interplay of their elements is a key problem in computational biology. Modern high-throughput multi-omics techniques now enable access to each regulatory network level, even at single-cell resolution (Lee *et al.*, 2020; Li *et al.*, 2021).

However, combining multi-omic measurements and reconstructing the underlying regulatory network remains challenging (Hawe *et al.*, 2019). Generally, sparse interaction networks in the form of directed or undirected graphs are constructed from dynamic interventional omics data or large observational data using different classes of statistical methods (Hawe *et al.*, 2019). Common approaches are either correlation-based (Langfelder and Horvath, 2008), use techniques from information theory (Song *et al.*, 2012; Margolin *et al.*, 2006; Lachmann *et al.*, 2016), or use (regularized) regression and variable selection frameworks to infer graphical models (Krumsiek *et al.*, 2011; Schäfer and Strimmer, 2005; Petralia *et al.*, 2015). Most recent methods also integrate prior knowledge (Sass *et al.*, 2013; Li and Jackson, 2015), such as experimentally

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.14.488153; this version posted April 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

determined protein-protein interaction networks, known metabolic pathways, or even predicted miRNA-mRNA interactions (List *et al.*, 2019). One such recent approach is KiMONo, **K**nowledge gu**I**ded **M**ulti-**O**mics **N**etw**o**rk inference approach (Ogris *et al.*, 2021). KiMONo is a two-step prior knowledge-based approach for multi-omic regulatory network inference. In the first step, the framework uses the whole dataset to model each omic element individually, detecting statistical effects between them. Here, the inference complexity is decreased by pre-selecting feature dependencies based on existing prior knowledge of biological mechanisms. The framework combines all models in a second step, assembling a multi-omic graph with the input features as nodes linked via edges representing the detected effects.

A major drawback of most network inference methods is their inability to handle missing data. It is often necessary to combine multiple studies that only partially measure the same omics levels to reach sample sizes adequate for network inference, creating patterns of block-wise missingness. Many classical regulatory network inference methods ignore missing data and focus only on analyzing complete cases, thus underutilizing the collected data set and severely limiting the amount of information used. Removing samples with missing features can also lead to biased estimates if the missingness is not completely random (Rubin, 2004), potentially affecting the extracted regulatory network. Multiple imputation (Donders *et al.*, 2006) is another popular approach to deal with missingness, followed by applying any classical network inference method using *ad hoc* rules to harmonize variable selection across multiply-imputed datasets (Wood *et al.*, 2008). However, Ganti and Willet demonstrated that such two-step approaches can be sub-optimal (Ganti and Willett, 2015) and instead require integrated or more general frameworks to handle missing data and variable selection jointly.

In recent years, advances have been made in using sparse graphical models for data with missing information. These approaches can be roughly categorized in Bayesian methods using data augmentation strategies (Ibrahim *et al.*, 2008), methods using pooled posterior (Yang *et al.*, 2005), or bootstrapped inclusion probabilities (Heymans *et al.*, 2007; Liu *et al.*, 2016), methods performing variable selection through stacked (Wan *et al.*, 2015; Wood *et al.*, 2008) or group Lasso integrated multiple imputation methods (Chen and Wang, 2013; Geronimi and Saporta, 2017; Marino *et al.*, 2017; Du *et al.*, 2022), low-rank matrix completion (Choi and Tibshirani, 2013; Ganti and Willett, 2015), inverse probability weighting (Johnson *et al.*, 2008), Lasso regularized inverse covariance estimation (Loh and Wainwright, 2011; Städler and Bühlmann, 2012; Takada *et al.*, 2018; Datta and Zou, 2017), and Expectation-Maximization-based approaches (Shen and Chen, 2012; Sabbe *et al.*, 2013). While most methods address the missingness of individual features, some methods exist that explicitly model block-missingness (Yu *et al.*, 2020; Xue and Qu, 2021; Du *et al.*, 2022; Gentry *et al.*, 2021).

Incorporating such approaches in multi-omics network inference is attractive. It extends the application of tools such as KiMONo to omic types such as metabolomics and proteomics, where missing features occur frequently, or to single-cell RNA sequencing data where missingness is inevitable due to stochastic gene expression and low capture efficiency. However, a comprehensive benchmark of existing methods that can handle missing data is lacking. We, therefore, extended the KiMONo framework with various regression-based approaches that integrate and combine prior imputed data (Du *et al.*, 2022) and Lasso regularized inverse covariance estimation methods (Takada *et al.*, 2018; Datta and Zou, 2017). We systematically evaluated how these methods can handle gradually increasing levels of artificial noise and missing and block-missing information for regulatory network inference on single- and multi-omics data. We evaluate the method robustness per regression model via the root mean squared error (RMSE) and R^2 and compare each method against method-specific baseline networks inferred without missing information or noise with precision, recall, and the F1 measure. In addition, we compare each perturbation-based network against an original KiMONo inferred network to assess how robust these networks are across methods. Finally, we also compare against networks inferred with the original KiMONo method after k-nearest neighbor imputation to assess the performance gain of methods that handle missingness implicitly.

We observed that approaches explicitly handling missingness in a two-step manner performed best over a wide range of random, block-missingness, and noise levels while implicit methods performed worst and were generally unstable.

2 Methods

2.1 Regression-based Methods for Network Inference and Imputation.

We focused on methods with a working R implementation and consistent documentation. These requirements left us with five advanced statistical approaches of three categories (1) stacked and (2) grouped multiple-imputation, as well as (3) Lasso-based inverse covariance estimation approaches (Table 1, detailed description see Supplementary Information S1). All mentioned methods have been integrated into the KiMONo framework. Source code of KiMONo can be found on GitHub (https://github.com/cellmapslab/kimono), while detailed benchmarking results and code can be found at Zenodo (Henao *et al.*, 2022) (https://doi.org/10.5281/zenodo.6450228).

Table 1*: Inference models included in this benchmark and capable of dealing with missing data.*

Method	Category	Citation
knnKiMONo	single imputation + KiMONo	(Troyanskaya *et al.*, 2001); (Ogris *et al.*, 2021)
SALasso	stacked multiple imputation	(Du *et al.*, 2022)
GALasso	grouped multiple imputation	(Du *et al.*, 2022)
HMLasso	inverse covariance estimation	(Takada *et al.*, 2018)
CoCoLasso	inverse covariance estimation	(Datta and Zou, 2017); (Takada *et al.*, 2018)
BDCoCoLasso	inverse covariance estimation	(Escribe *et al.*, 2021)

kNN-imputation & KiMONo (knnKiMONo): We implemented a two-step approach that firstly imputes missing information using nearest neighbor averaging followed by applying the classical KiMONo. The kNN-based imputation method (Troyanskaya *et al.*, 2001) implemented in the R package impute v1.46.0 was applied separately to individual omics layers and other covariates. Originally designed for the imputation of gene expression data, the method replaces missing values by averaging non-missing values of its nearest neighbors. If the percentage of missing data allowed for every variable, e.g., a single gene exceeds 50% (default), the missing values are imputed using the overall mean per sample. Only samples with missingness less than 80% (default) were considered for the imputation. Further, algorithm's parameters were set to default values: the number of neighbors used in the imputation was set to k=10, and the largest block of variables imputed using the kNN algorithm before recursively dividing the feature into smaller chunks was set to max = 1500.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.14.488153; this version posted April 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

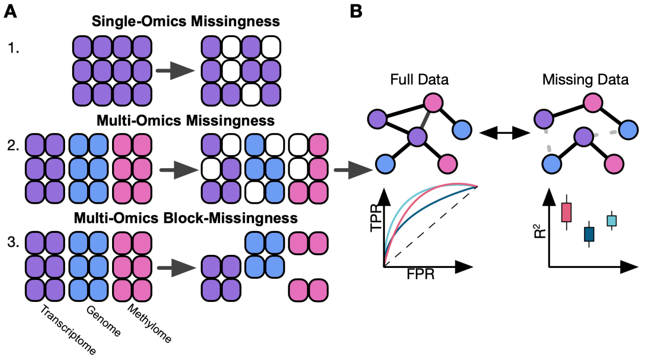

Figure 1: *Benchmark schematic.* ***A)*** *Three common missingness scenarios in regulatory network inferences are tested: 1) single-omics, 2) multi-omics random missingness of individual elements, and 3) block-wise missingness in which entire omics layers are missing for an individual.* ***B)*** *We tested five approaches of two categories: 1) Two-step approaches that first impute and then aggregate imputation through, and 2) Inverse covariance estimation approaches that implicitly handle missingness during inference. We inferred regulatory networks from full data and data missing for each method and compared the resulting networks with multiple performance metrics.*

Stacked Adaptive Lasso (SALasso): Stacked approaches combine prior D-times multiply imputed datasets by averaging over them during inference, making such approaches applicable to existing sparse regression framework. We use the SALasso implementation released in the R package miselect 0.9.0 (Du *et al.*, 2022). We tested 50 values with a lambda.min.ratio of 1e-4 in a 5-fold cross-validation with and without adaptive weights, while sample weights were set to be uniformly distributed.

Grouped Adaptive Lasso (GALasso): Similar to SALasso, GALasso pools prior imputed datasets by adding a group LASSO penalty term enforcing consistent variable selection across multiply imputed datasets. We use the GALasso implementation released in the R package miselect 0.9.0 (Du *et al.*, 2022). We tested 50 lambda values with a lambda.min.ratio of 1e-4 in a 5-fold cross-validation with and without adaptive weights, while the sample weights were set to be uniformly distributed.

Convex Conditioned Lasso (CoCoLasso): CoCoLasso is an inverse covariance estimation method for high-dimensional data with missing values. The main idea is to reformulate the Lasso regression by working with the sample covariance matrix of X, $S = \frac{1}{n}X'X$, and the sample covariance vector of X and y, $\rho = \frac{1}{n}X'y$. With this reformulation, β is estimated via S and ρ instead of X and y. We use the CoCoLasso implementation released in the R package HMLasso 0.0.1 (Takada, M., Fujisawa, H., & Nishikawa, T., 2019) with the following selection of hyperparameters: For lambda, we tested 50 values with a lambda.min.ratio of 1e-1 in a 5-fold cross-validation.

Lasso with High Missing rate (HMLasso): HMLasso can be seen as an optimally weighted modification of CoCoLasso according to the missing ratio. HMLasso uses the mean imputation method. Instead of X the mean imputed data variable, Z is used, where $Z_{jk} = X_{jk}$ for an observed element and $Z_{jk} = 0$ otherwise. We use the HMLasso implementation released in the R package HMLasso 0.0.1 (Takada, M., Fujisawa, H., & Nishikawa, T., 2019) with the following selection of hyperparameters: For α, we tested values between 0.5 and 2 with an interval of 0.5. For lambda, we tested 50 values with a lambda.min.ratio of 1e-1 in a 5-fold cross-validation.

Block-descent-CoCoLasso (BDCoCoLasso): To improve the computational efficiency of CoCoLasso, BDCoCoLasso implements a block coordinate descent strategy (Escribe *et al.*, 2021), where it projects the covariance matrix onto a positive semidefinite subspace on the corrupted subblocks. Unlike uncorrupted covariates that are measured without any error, corrupted covariates are measured with an error or are missing, leading to inconsistent estimates. The covariance matrix X_{nxp} is separated into $[X_{1_{nxp_2}}, Z_{2_{nxp_2}}]$ where $X_{nxp_1}$and $Z_{2_{nxp_2}}$ corresponds to the uncorrupted and corrupted covariates, respectively. Then, β is defined as $\beta = (\beta_1, \beta_2)$ where β_1 and β_2 correspond to the coefficient vector for the uncorrupted and corrupted covariates, respectively. We use the BDCoCo implementation released in the R package BDCoCoLasso v0.0.0.9000 (https://github.com/celiaescribe/BDCoCoLasso) with the following selection of hyperparameters: For lambda, we tested 50 values within a range between 0 and 1e-2 with an adaptive cross-validation schema.

2.2 Datasets

We collected triple-omics data (transcriptome, copy number variation (CNV), and methylation data) as well as clinical data from the breast invasive carcinoma atlas, which is one of the most comprehensive multi-omic data resources to date with 871 matched samples, from the PanCancer Projects (Weinstein *et al.*, 2013) using The Cancer Genome Atlas (TCGA) data portal and the cBioPortal (Gao *et al.*, 2016) (retrieved on 03/07/2022). All samples containing missing information were removed to construct a complete data set as the baseline, thus restricting the data sets to 604 patients. Similarly, features with low variance were removed, resulting in 11,530 transcriptomics features, 1,366 methylation features, and 84 copy number variation (CNV) features.

2.3 Prior network generation

We extracted protein-protein interactions from the BioGrid interactome (Release 3.5.188) (Oughtred *et al.*, 2021), associated these interactions with the extracted gene expression information, and linked each CNV and methylation site to its associated gene since both omics layers were already annotated to gene identifiers. The final prior network contained 11,645 nodes (10,848 genes, 84 CNVs, and 713 methylation sites).

2.4 Network-based multiple imputation

Stacked and grouped adaptive Lasso approaches require multiple imputed data as input. Nevertheless, multiple imputation methods do not scale well to high dimensional data with high missingness, and standard implementations such as those offered in the R package MICE, take multiple hours to days to finish. Thus, we developed a novel network-guided multiple imputation by chained equation approach (ngMICE) by utilizing KiMONo's prior network. Instead of considering all covariates for imputation, we restrict each imputation attempt to the covariates that are directly linked to the missing covariate in a prior network as other covariates will be removed during network inference by KiMONo and therefore can be neglected. The number of covariates can be further reduced by correlation-based filtering. For missing elements retaining less than k covariates for imputation, the top k correlated covariates are used. Once the covariate matrix has been constructed as described, the standard MICE procedure is run. We used the R package MICE 3.14.0 (van Buuren and Groothuis-Oudshoorn, 2011), with Bayesian linear regression as a multiple imputation approach and an absolute Pearson correlation coefficient of 0.1 as the threshold.

ngMICE performed similarly to kNN-based imputation in terms of RMSE across omics types and missingness with slightly worse average performance (Supplementary Figure 1).

2.5 Benchmark

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.14.488153; this version posted April 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

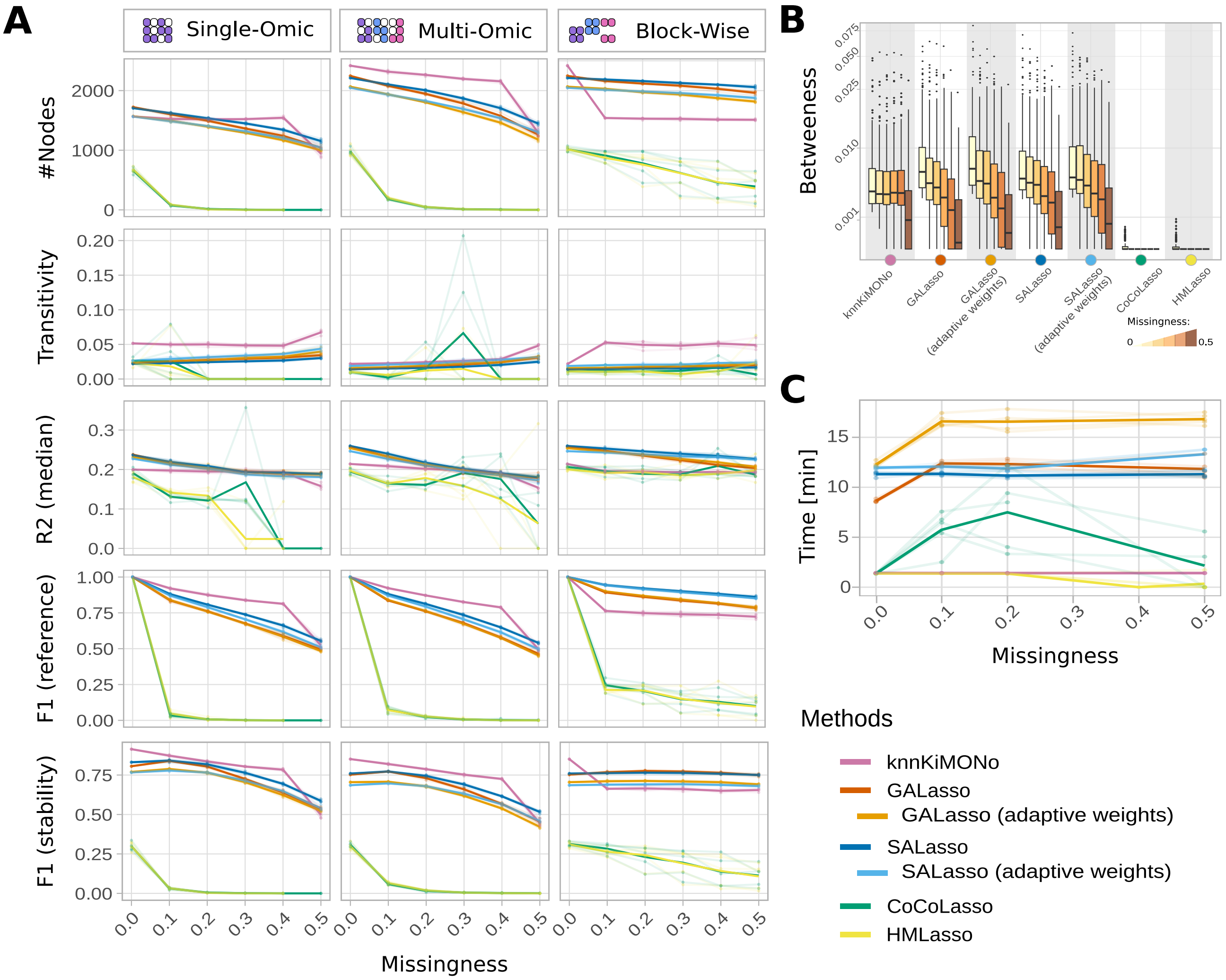

Figure2: Benchmark results across all experimental setups. Transparent lines denote individual runs, while bold lines refer to the average performance. ***A)*** *Performance under Single-Omic, Multi-Omic, and Block-missingness were evaluated using network size (number of nodes), transitivity (global clustering coefficient), median R^2, and F1 scores compared to a reference (i.e., the same method applied on the full data and stability selected networks generated with KiMONo).* ***B)*** *Illustrating the topological change with increased missingness based on the 200 network nodes with the highest betweenness centrality.* ***C)*** *Comparison of computing time.*

We assess the performance of the selected inference methods in the presence of missing data via simulating two typical scenarios - (1) random missing information in a single omic level and across multiple levels, as well as (2) a block-wise missing structure, i.e., data in one or more omic levels may be unobserved (Figure 1a). To stress the method's capabilities even further, we decreased the signal-to-noise ratio by systematically adding covariate-specific white noise to the input data. Each experiment was repeated 5 times for robust performance estimation and corrected for confounding age and sex effects.

Single-Omics missing: We selected the transcriptomics level as a single-omic layer to test the different models' capabilities to infer gene regulatory networks with less directly informative co-correlation structures that could be used to impute the missing gene expression information. We then removed $m \in \{0\%, 10\%, 20\%, 30\%, 40\%, 50\%\}$ randomly selected entries from the input data. Additionally, we added white noise with increasing intensity to the data by drawing from a normal distribution $\epsilon \sim N(0, a\sigma_g^2)$ per gene with a gene-specific variance term estimated from the real data and $a \in \{0, 0.5, 1.5\}$.

Multi-Omics missing: To test the models' capabilities to handle more complex co-correlation structures that could potentially be exploited for better imputation, we expanded the single-omics experiment to jointly consider the three available omics types. As before, we randomly removed $m \in \{0\%, 10\%, 20\%, 30\%, 40\%, 50\%\}$ of entries independently per omics layer and added feature-dependent white noise to the data as described before while ensuring to bound the beta values of the methylation data to the range between 0 and 1.

Multi-Omics Block-Missing: To test the capabilities of the method to handle block-wise missing information, in which an entire omics layer is missing for a random selection of patients, we removed $m \in \{0\%, 10\%, 20\%, 30\%, 40\%, 50\%\}$ patients per omics layer such that at least two omics-layers still remained per individual. Additionally, we added white noise to the remaining samples as described before.

Downsampling: Similarly, to identify the minimal number of samples required to infer reliably the regulatory network, we downsampled the dataset to $k \in \{90\%, 80\%, 70\%, 60\%, 50\%\}$ of samples.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.14.488153; this version posted April 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Runtime: We tested the runtime of each method on a dedicated machine with an AMD EPYC 7502P 32-Core Processor with 2.5GHz base clock speed and 859.4 RAM using the multi-omics missingness experiment with the same configurations as before. We ran all experiments with ncores=60 (with hyperthreading enabled).

2.4 Evaluation Metrics

To construct the final network from the individual regressions, we applied a strict filter connecting independent to dependent variables if their beta coefficient was non-zero and their $R^2 > 0.1$.

Prediction Metrics: To measure the prediction qualities of each regression model, we recorded the root means squared error (RMSE) and R^2 respectively, and compared their distribution based on a Wilcoxon rank-sum test.

Network Reconstruction Metrics: To measure the methods' abilities to handle missing data well, we compared the inferred regulatory networks from missing data to their counterpart inferred on full data, respectively, and calculated precision, recall, and F1 of the recovered network edges. Similarly, we inferred a ground truth network with KiMONo using stability selection by repeating network inference *k* times on different data splits and averaging over the resulting coefficients and R^2 values before constructing the network, improving the robustness of the graph. The final stability-selected and conservatively filtered network consisted of 2,458 nodes (2,182 genes, 63 CNV, 211 methylations) and 6,554 edges. We compared the stability-selected reference network to all inferred networks on missing data to determine performance differences across the individual methods.

Topological Metrics: The interpretation of complex heterogeneous networks and identifying key modules and important network nodes relies on the topological network features. Hence, it is also vital to evaluate if the methods can robustly infer topological structures. Therefore, we use multiple network-based metrics such as node-degree distribution, betweenness centrality, and clustering coefficient to quantify and compare the topological changes of the networks inferred from missing data. Node degree indicates the sparseness of the network, while betweenness centrality indicates how interconnected the network is, and the global clustering coefficient (transitivity) indicates how densely connected neighboring nodes are.

3 Results

For BDCoCoLasso, we could not use the inferred model on incomplete samples and could not calculate RMSE and R² metrics for a significant proportion of edges. Due to this reason, we could not infer complete networks on data with missing values (Supplementary Information S1). We, therefore, refrain from discussing the performance of BDCoCoLasso in the following discussion.

3.1 Most topological features can be conserved in data with missingness

One aspect of robustness in presence of noise or missingness is that the network topology should remain largely unchanged. To investigate this, we computed multiple network properties across all benchmarking scenarios (Figure 2A&B).

Our benchmark showed that induced missingness increases the transitivity while decreasing the number of nodes and edges regarding the full and uncorrupted data topological network properties. The largest heterogeneous networks were produced by knnKiMONo (~2400 nodes) on the full dataset (Figure 2A). This was followed by GALasso and SALasso modeling on average over 2,200 nodes, CoCoLasso (~950 nodes), and HMLasso (~900 nodes) (Supplementary Figure 2). While all methods suffered under high missingness resulting in smaller networks, knnKiMONo appeared to be robust in low to medium missingness conditions. As the networks fell apart into unconnected modules, the top 200 nodes' average betweenness of centrality decreased while the global transitivity increased (Figure 2A&B). Similar patterns could be observed through the number of inter and intra omic edges. Here, the constant loss of edges showed a stable ratio indicating no bias towards a specific omic layer. Briefly, across all benchmark settings, the methods showed a decreasing network size with increased missingness. GALasso and SALasso were less affected in terms of noise, while knnKiMONo produces the most robust results in terms of missingness (Supplementary Figure 2).

3.2 kNN imputation-based models perform best for single- and multi-omics data with random missingness

In both the single- and multi-omics setting, where we randomly perturbed omics layers independently, knnKiMONo was the best performing method reaching F1 scores of 0.921±0.005 on in the single omics and 0.923±0.002 on the multi-omics data at 10% missingness followed by SALasso (single-omics: 0.880±0.005, multi-omics: 0.882±0.004) and GALasso (single-omics: 0.836±0.009, multi-omics: 0.837±0.003) when compared to a reference network inferred on full data (Figure 2A; Supplementary Figure 3&4). The performance gradually decreased with increasing missingness and noise levels to a similar degree for all three methods dropping to 0.348±0.018 and 0.315±0.011 for knnKiMONo, 0.381±0.009 and 0.339±0.013 for SALasso, as well as 0.318±0.015 and 0.272±0.008 for GALasso, for single- and multi-omics data, respectively, at 50% missingness and a medium noise (a=0.5). At higher noise levels, the performance of all three methods declined dramatically reaching only F1 scores below 0.1. HMLasso and CoCoLasso came in last, reaching F1 scores below 0.1 both on single- and multi-omics data even for samples with only 10% missingness and no noise (Supplementary Figure 4). Both methods even failed to infer networks in 11/15 and 11/15 single-cell experiments at 30% and 40% missingness, as well as in 9/11 and 8/11 multi-omics experiments, at 40% and 50% missingness, respectively.

Similar behavior could be observed when comparing the inferred networks to the stability selection-based reference network: knnKiMONo performed best with an F1 score of 0.873±0.005 and 0.820±0.004 at 10% missingness and no noise on single- and multi-omics data respectively, followed by SALasso (single-omcis: 0.841±0.007, multi-omics: 0.772±0.003) and GALasso (single-omics: 0.840±0.005, multi-omics: 0.772±0.001) (Figure 2A; Supplementary Figure 5). SALasso did marginally outperform knnKiMONo in samples with high noise and/or high missingness. At high noise levels (a=1.5), even in the absence of missingness, none of the methods exceeded F1 scores of 0.1 for both single- and multi-omics data, with the exception of SALasso with 0.113±0.013 for single-omics, and 0.121±0.002 for multi-omics (Supplementary Figure 5).

For both SALasso and GALasso adaptive weights had a marginal impact on performance with a slight negative effect. Only at high noise levels did adaptive weights stabilize performance (Supplementary Figure 4&5).

3.4 SALasso performs best for data with block-missingness

When investigating block-missingness, i.e., where entire omics layers are missing for a subset of patients, SALasso performed overall best, outperforming knnKiMONo and GALasso (Figure 2A; Supplementary

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.14.488153; this version posted April 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Figure 3). The SALasso inferred networks were consistent with the networks inferred on full data across different block missingness levels, with SALasso reaching F1 scores between 0.945±0.004 at 10% missingness (a=0) and 0.688±0.002 at 50% missingness (a=0.5). With higher noise, the performance dropped below an F1 score of 0.1 (Supplementary Figure 4). This behavior could be observed for GALasso and knnKiMONo, although with generally lower F1 scores. Notably, knnKiMONos networks almost exclusively consisted of gene nodes, while all other methods had a proportional representation of all omics types. The implementation of kNN-imputation used here is not able to handle entire block-missing samples and, consequently, knnKiMONo removes a substantial amount of the features. The explanation is the sample missing removal takes the dependent vector as reference. However, the sample missing in this vector differs from the remain omics layers provoking the deletion of those missing sample from the same omic layer this vector was extracted, but conserving the sample missingness from the rest of omics matrices.

Similar behavior could be observed when comparing the inferred networks to the stability selection inferred network. Both SALasso and GALasso outperformed knnKiMONo with SALasso overall reaching the highest F1 scores of 0.763±0.003 (10% missingness, a=0) to 0.655±0.007 (50% missingness, a=0.5), while knnKiMONo reached F1 scores of 0.663±0.003 (10% missingness, a=0) to 0.472±0.011 (50% missingness, a=0.5). High noise reduced the F1 scores below 0.1. Adaptive weights had negligible effects and only improved performance markedly at high noise levels (a=1.5) (Supplementary Figure 5).

HMLasso and CoCoLasso performed worst, reaching average F1 scores of 0.212±0.022, and 0.246±0.039 at 10% block-missingness, dropping to 0.063±0.033 and 0.073±0.031 respectively with 50% block-missingness and medium noise (a=0.5) (Supplementary Figure 4). Comparing those methods to the stability-selection-based reference depicted a similar picture. Average F1 scores of 0.264±0.027 and 0.283±0.035 could only be reached at 10% block-missingness, respectively, while higher missingness and noise levels led to F1 scores below 0.05 (Supplementary Figure 5).

3.5 knnKiMONo and SALasso are least affected by sample size reduction

All methods showed a decrease in concordance already at 90% of the total sample size (Supplementary Figure 6&7). HMLasso and CoCoLasso were most severely affected, with F1 scores at 0.222±0.015 and 0.205±0.021, albeit remaining at this performance level for larger sample size reductions (Figure 2A; Supplementary Figure 6). knnKiMONo reached an F1 score of 0.918±0.020 and slowly decreased in performance with increasing sample size reduction to 0.820±0.020 at 50% sample size. SALasso and GALasso behaved similarly, reaching F1 scores of 0.908±0.004 and 0.824±0.009 respectively at 90% sample size, and GALasso gradually decreased with smaller sizes, reaching 0.864±0.003 and 0.793±0.009 at 50% sample size. At these high sample reductions (40 and 50%), SALasso is even more stable than knnKiMONo. Adaptive weights, as before, reduced performance marginally (Supplementary Figure 6&7).

3.6 HMLasso is the fastest approach in datasets with no to medium missingness

knnKiMONo, HMLasso, and CoCoLasso had a similar runtime of 82 - 85 sec on the full dataset without missing information and gradually increased in runtime with rising missingness (Figure 2C). HMLasso was the fastest approach with an average runtime of 81.843±0.375 sec at low to medium missingness levels. Only in scenarios with high missingness HMLasso was outperformed by knnKiMONo, reaching an average runtime of 97.089±1.327 sec. knnKiMONo demonstrated a very consistent runtime across all missingness levels with an average runtime of 97.725±7.995 sec. CoCoLasso behaved similarly but was affected by the degree of missingness, reaching maximum average runtimes of 449.6±225.2 sec. GALasso and SALasso were the slowest, with average runtimes of 739±15.532 sec and 919.785±11.833 sec on the complete dataset, respectively, of which 241.191±2.975 sec were dedicated to imputation. Overall runtime gradually increased for both methods reaching an overall average runtime of 1148.871±245.077 sec and 1184.688±157.794, respectively of which on average 507.674±58.126 sec was spent in imputation. For GALasso, adaptive weights calculations added another 275.363±54.180 sec on average to the overall runtime.

4 Conclusion

Due to economic or technical restrictions, missingness of individual values or block-missingness of entire omics layers in a subset of samples is typical for high-throughput multi-omics experiments, rendering multi-omics network inference challenging. In particular, single-cell experiments, which have become increasingly prevalent and offer unprecedented insights into the molecular landscape, are affected by sparsity and missingness. In this study, we benchmarked novel regression approaches that can handle missing information across common missingness scenarios in single and multi-omics experiments and integrated these approaches into KiMONo, a recent approach for network-guided multi-omics network inference.

Generally, we observed that approaches explicitly handling missing information through prior imputation overwhelmingly outperform methods that implicitly handle missingness. Specifically, kNN-imputation combined with the standard KiMONo approach performed best, closely followed by SALasso and GALasso approaches that combine multiple-imputation results. Both SALasso and GALasso probably suffered from the lower imputation quality of the network-based multiple-imputation approach (ngMICE) we applied, propagating the imputation uncertainty into network inference. Multiple-imputation in high-dimensional data is generally a challenge since existing approaches do not scale to the number of covariates typically encountered in multi-omics studies. Here, dimensionality reduction methods for multiple imputation (Hodge *et al.*, 2019), latent factor models (Argelaguet *et al.*, 2020), or deep-learning-based approaches (Qiu *et al.*, 2020; Gayoso *et al.*, 2021; Lotfollahi *et al.*, 2022) might improve multiple imputation and therfore network inference quality even over the simple kNN-imputation-based approach. The potential superior performance of multiple-imputation-based approaches was indicated by SALasso already outperforming knnKiMONo in block-missingness cases and in high missingness-high noise multi-omics cases. A benefit of such explicit approaches is their ability to use and adequately address prior imputed datasets often provided by larger consortia.

Surprisingly, implicit methods relying on inverse covariance matrix estimation performed poorly, indicating that more research is needed to make these approaches robust.

We note that a true gold standard for evaluating the performance of network inference methods is missing. In its absence, we rely on a network inferred from complete, unperturbed data that nevertheless is likely to contain both false positive and false negative interactions which may affect the results. Hence, our reference network is not suited for evaluating methods following different principles for inference as GENIE3 (Huynh-Thu *et al.*, 2010), or other non-linear approaches. However, a comparison with derivatives of the Lasso method is reasonable, as differences in the results can be attributed to each method's ability to handle missingness or noise.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.14.488153; this version posted April 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

In summary, we found explicit methods to be more robust than methods implicitly handling missingness. While most methods were tolerant to high levels of missingness, they were strongly affected by noise. While HMLasso was the fastest tested method, knnKiMONo showed the best tradeoff between performance and runtime and is thus our recommended approach for handling missingness in KiMONo. While we see room for further method improvements, particularly with respect to multiple imputations of high-dimensional data and the robustness of inverse covariance methods, our results show that robust multi-omics network inference with KiMONo is feasible and thus allows users to leverage available multi-omics data to its full extent.

Contributions

JDH, MiL, MA, AG implemented the different lasso approaches. **JDH** implemented the benchmark framework**. JDH, BS** analyzed the results**. BS, CO, MaL** conceived and supervised the study. **JDH, BS, CO, MaL** wrote the original manuscript. **MaL, CO, BS** reviewed and edited the manuscript. All authors read and approved the final manuscript.

Funding

This work has been supported by the German Center of Lung Research (DZL). BS also acknowledges financial support from the Helmholtz International Lab “Causal Cell Dynamics”. MiL was supported by the Hanns Seidel Foundation. This study was also supported by the BMBF (German Federal Ministry of Education and Research) Project TRY-IBD Grant 01ZX1915B for CO and AG.

Conflict of Interest: none declared

References

Argelaguet,R. *et al.* (2020) MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol.*, **21**, 111.

van Buuren,S. and Groothuis-Oudshoorn,K. (2011) mice: Multivariate Imputation by Chained Equations in R. *J. Stat. Softw.*, **45**, 1–67.

Chen,Q. and Wang,S. (2013) Variable selection for multiply-imputed data with application to dioxin exposure study. *Statistics in Medicine*, **32**, 3646–3659.

Choi,Y. and Tibshirani,R. (2013) An Investigation of Methods for Handling Missing Data with Penalized Regression. *arXiv [stat.AP]*.

Datta,A. and Zou,H. (2017) CoCoLasso for high-dimensional error-in-variables regression. *aos*, **45**, 2400–2426.

Donders,A.R.T. *et al.* (2006) Review: A gentle introduction to imputation of missing values. *J. Clin. Epidemiol.*, **59**, 1087–1091.

Du,J. *et al.* (2022) Variable selection with multiply-imputed datasets: choosing between stacked and grouped methods. *J. Comput. Graph. Stat.*, 1–35.

Escribe,C. *et al.* (2021) Block coordinate descent algorithm improves variable selection and estimation in error-in-variables regression. *Genet. Epidemiol.*, **45**, 874–890.

Ganti,R. and Willett,R.M. (2015) Sparse Linear Regression With Missing Data. *arXiv [stat.ML]*.

Gao,J. *et al.* (2016) Abstract 5277: The cBioPortal for cancer genomics and its application in precision oncology. *Cancer Res.*, **76**, 5277–5277.

Gayoso,A. *et al.* (2021) Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nat. Methods*, **18**, 272–282.

Gentry,A.E. *et al.* (2021) Missingness Adapted Group Informed Clustered (MAGIC)-LASSO: A novel paradigm for prediction in data with widespread non-random missingness. *bioRxiv*, 2021.04.29.442057.

Geronimi,J. and Saporta,G. (2017) Variable selection for multiply-imputed data with penalized generalized estimating equations. *Comput. Stat. Data Anal.*, **110**, 103–114.

Hawe,J.S. *et al.* (2019) Inferring Interaction Networks From Multi-Omics Data. *Front. Genet.*, **10**, 535.

Henao,J.D. *et al.* (2022) Multi-Omics Regulatory Network Inference in the Presents of Missing Data.

Heymans,M.W. *et al.* (2007) Variable selection under multiple imputation using the bootstrap in a prognostic study. *BMC Med. Res. Methodol.*, **7**, 33.

Hodge,D.W. *et al.* (2019) Multiple imputation using dimension reduction techniques for high-dimensional data. *arXiv [stat.ME]*.

Huynh-Thu,V.A. *et al.* (2010) Inferring regulatory networks from expression data using tree-based methods. *PLoS One*, **5**.

Ibrahim,J.G. *et al.* (2008) Bayesian variable selection for the Cox regression model with missing covariates. *Lifetime Data Anal.*, **14**, 496–520.

Johnson,B.A. *et al.* (2008) Penalized Estimating Functions and Variable Selection in Semiparametric Regression Models. *J. Am. Stat. Assoc.*, **103**, 672–680.

Krumsiek,J. *et al.* (2011) Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Syst. Biol.*, **5**, 21.

Lachmann,A. *et al.* (2016) ARACNe-AP: gene network reverse engineering through adaptive partitioning inference of mutual information. *Bioinformatics*, **32**, 2233–2235.

Langfelder,P. and Horvath,S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, **9**, 559.

Lee,J. *et al.* (2020) Single-cell multiomics: technologies and data analysis methods. *Exp. Mol. Med.*, **52**, 1428–1442.

List,M. *et al.* (2019) Large-scale inference of competing endogenous RNA networks with sparse partial correlation. *Bioinformatics*, **35**, i596–i604.

Liu,Y. *et al.* (2016) VARIABLE SELECTION AND PREDICTION WITH INCOMPLETE HIGH-DIMENSIONAL DATA. *Ann. Appl. Stat.*, **10**, 418–450.

Li,Y. *et al.* (2021) Advances in bulk and single-cell multi-omics approaches for systems biology and precision medicine. *Brief. Bioinform.*, **22**.

Li,Y. and Jackson,S.A. (2015) Gene Network Reconstruction by Integration of Prior Biological Knowledge. *G3* , **5**, 1075–1079.

Loh,P.-L. and Wainwright,M.J. (2011) High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *Adv. Neural Inf. Process. Syst.*, **24**.

Lotfollahi,M. *et al.* (2022) Multigrate: single-cell multi-omic data integration. *bioRxiv*, 2022.03.16.484643.

Margolin,A.A. *et al.* (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, **7 Suppl 1**, S7.

Marino,M. *et al.* (2017) Covariate Selection for Multilevel Models with Missing Data. *Stat*, **6**, 31–46.

Ogris,C. *et al.* (2021) Versatile knowledge guided network inference method for prioritizing key regulatory factors in multi-omics data. *Sci. Rep.*, **11**, 6806.

Oughtred,R. *et al.* (2021) The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci.*, **30**, 187–200.

Petralia,F. *et al.* (2015) Integrative random forest for gene regulatory network inference. *Bioinformatics*, **31**, i197–205.

Qiu,Y.L. *et al.* (2020) Genomic data imputation with variational auto-encoders. *Gigascience*, **9**.

Romero,I.G. *et al.* (2012) Comparative studies of gene expression and the evolution of gene regulation. *Nat. Rev. Genet.*, **13**, 505–516.

Rubin,D.B. (2004) Multiple Imputation for Nonresponse in Surveys John Wiley & Sons.

Sabbe,N. *et al.* (2013) EMLasso: logistic lasso with missing data. *Stat. Med.*, **32**, 3143–3157.

Sass,S. *et al.* (2013) A modular framework for gene set analysis integrating multilevel omics data. *Nucleic Acids Res.*, **41**, 9622–9633.

Schäfer,J. and Strimmer,K. (2005) An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, **21**, 754–764.

Shen,C.-W. and Chen,Y.-H. (2012) Model selection for generalized estimating equations accommodating dropout missingness. *Biometrics*, **68**, 1046–1054.

Song,L. *et al.* (2012) Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC Bioinformatics*, **13**, 328.

Städler,N. and Bühlmann,P. (2012) Missing values: sparse inverse covariance estimation and an extension to sparse regression. *Stat. Comput.*, **22**, 219–235.

Takada,M. *et al.* (2018) HMLasso: Lasso with High Missing Rate. *arXiv [stat.ML]*.

Troyanskaya,O. *et al.* (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520–525.

Wan,Y. *et al.* (2015) Variable selection models based on multiple imputation with an application for predicting median effective dose and maximum effect. *J. Stat. Comput. Simul.*, **85**, 1902–1916.

Weinstein,J.N. *et al.* (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.*, **45**, 1113–1120.

Wood,A.M. *et al.* (2008) How should variable selection be performed with multiply imputed data? *Stat. Med.*, **27**, 3227–3246.

Xue,F. and Qu,A. (2021) Integrating Multisource Block-Wise Missing Data in Model Selection. *J. Am. Stat. Assoc.*, **116**, 1914–1927.

Yang,X. *et al.* (2005) Imputation and variable selection in linear regression models with missing covariates. *Biometrics*, **61**, 498–506.

Yu,G. *et al.* (2020) Optimal Sparse Linear Prediction for Block-missing Multi-modality Data without Imputation. *J. Am. Stat. Assoc.*, **115**, 1406–1419.