MCNET: Multi-Omics Integration for Gene Regulatory Network Inference from scRNA-seq

Ansh Tiwari[1] and Sachin Trankatwar[2]

[1,2] *Birla Institute of Technology & Science Pilani, Hyderabad Campus, Hyderabad, Telangana, 500078, India*
anshtiwari9899@gmail.com
strankatwar@gmail.com

ABSTRACT

Deep learning has emerged as a powerful approach in various domains, including biological network analysis. This paper investigates the advancements in computational techniques for inferring gene regulatory networks (GRNs) and introduces MCNET, a state-of-the-art deep learning algorithm. MCNET integrates multi-omics data to infer GRNs and extract biologically significant representations from single-cell RNA sequencing (scRNA-seq) data. By incorporating attention mechanisms and graph convolutional networks, MCNET captures intricate regulatory relationships among genes. Extensive benchmarking on diverse scRNA-seq datasets demonstrates MCNET's superiority over existing methods in GRN inference, scRNA-seq data visualization, clustering, and simulation. Notably, MCNET accurately predicts gene regulations on cell-type marker genes in the mouse cortex, validated by epigenetic data. The introduction of MCNET paves the way for advanced analysis of scRNA-seq data and provides a powerful tool for inferring GRNs in a multi-omics context. Moreover, this paper addresses the integration of multi-omics data in gene regulatory network inference, proposing MCNET as a method that efficiently analyzes and visualizes homogeneous gene regulatory networks derived from diverse omics data. The inference capability of MCNET is evaluated through extensive experiments with simulation data and applied to analyze the biological network of psychiatric disorders using human brain data.

1. INTRODUCTION

Understanding many biological processes requires knowledge not only about the biological entities themselves but also the relationships among them. For example, processes such as cell differentiation depend not only on which proteins are present, but also on which proteins bind together. A natural way to represent such processes is as a graph, also called a network, since a graph can model both entities as well as their interactions. Recent advances in experimental high-throughput technology have vastly increased the data output from interaction screens at a lower cost and resulted in a large amount of such biological network data (Reuter et al., 2015). The availability of this data makes it possible to use biological network analysis to tackle many exciting challenges in bioinformatics, such as predicting the function of a new protein based on its structure or anticipating how a new drug will interact with biological pathways. This wealth of new data, combined with the recent advances in computing technology that has enabled the fast processing of such data (Goodfellow et al., 2016), has reignited interest in neural networks (Sokolov et al., 2015; Parker, 1985; LeCun, 1985; Rumelhart et al., 1986) which date back to the 1970s and 1980s, and set the stage for the emergence of deep neural networks, a.k.a deep learning, as a new way to address these unsolved problems.

Biological systems on different levels of organization, from organelles and single cells to tissues, organs and entire organisms, constantly sense the environment and modulate their behavior to ensure optimal performance and fitness (López-Barneo et al., 2001; Rolland et al., 2006; Veal et al., 2007). The sensing of the environment is accomplished via numerous molecular mechanisms which ultimately result in coordinate activation and suppression of, often multiple, regulatory cascades affecting different and mutually dependent cellular processes. By propagating the perceived signal, the expression levels of genes coding for transcription factors (TFs) are adequately altered, leading to changes in the levels of transcripts encoding enzymatic proteins which affect metabolism and organism's tasks (Jacob & Monod, 1961).

Therefore, accurate reconstruction of the complete set of regulatory interactions, forming gene regulatory networks, is one of the key tasks in systems biology (Karlebach & Shamir, 2008). Gene regulatory networks have an important role in every process of life, including cell differentiation, metabolism, the cell cycle and signal transduction. By understanding the dynamics of these networks we can shed light on the mechanisms of diseases that occur when these cellular processes are

dysregulated. Accurate prediction of the behaviour of regulatory networks will also speed up biotechnological projects, as such predictions are quicker and cheaper than lab experiments Karlebach & Shamir (2008).

In biological systems, the gene regulatory interactions are transitory, as they depend on different factors, including: developmental, environmental, as well as internal, given by the genetic make-up of the organism. the expression of cognate genes is integrated in layers of iterative regulatory networks that ensure the performance not only of the whole cell, but also of the bacterial population, and even of the entire microbial community, in a changing environment (Cases & de Lorenzo, 2005).

High-throughput technologies for simultaneous measurement of gene expression have been used to capture the transitory behavior of thousands of genes upon internal and external perturbation in different biological systems, from bacteria and yeast to algae, plants and animals (Schulze & Downward, 2001; Blencowe et al., 2009; Rehrauer et al., 2009). The gathered gene expression levels reflect the underlying regulatory relationships, and, thus, can readily be used to reconstruct the operational regulatory networks.

With the increasing number of performed time-series experiments methods are needed to extract gene regulatory networks supported by all gathered data sets simultaneously. These experiments are over different time domains, with different sampling frequency under various conditions and conducted in different laboratories, which may affect the success of network reconstruction (Sima et al., 2009). In addition, each of these experiments is usually accompanied by a corresponding reference control experiment, whose profiles are used to determine differential gene behaviors (García de la Nava et al., 2004; Rapaport et al., 2013).

Reconstruction of gene regulatory networks is a classical problem in computational systems biology and various methods based on different sets of assumptions and applicable on data from particular experiments have been proposed, critically assessed and systematically reviewed (Hempel et al., 2011; Marbach et al., 2012; Omony, 2014).

In general, inference of gene regulatory networks begins with application of a similarity measure of choice on the investigated data set, resulting in a square similarity matrix. This similarity matrix can be sparsified by retaining only the values which are statistically significant after multiple hypothesis testing. A number of computational models (Rumelhart et al., 1986; Huynh-Thu et al., 2010; Chan et al., 2017; Matsumoto et al., 2017; Papili Gao et al., 2018) have attempted to incorporate GRN inference into their single-cell data analysis models. Current methods solely based on single-cell RNA sequencing (scRNA-seq) data also have explicit limitations. For example, it is common for GRN inference algorithms to use statistics algorithms that focus on the co-expression networks instead of decoding the casual relationships among TFs and their corresponding target genes (Chan et al., 2017; S. Kim, 2015). A number of computational models (Rumelhart et al., 1986; Huynh-Thu et al., 2010; Chan et al., 2017; Matsumoto et al., 2017; Papili Gao et al., 2018; Moerman et al., 2019) have attempted to incorporate GRN inference into their single-cell data analysis models.(Matsumoto et al., 2017; Papili Gao et al., 2018) or tree-based models (Huynh-Thu et al., 2010; Moerman et al., 2019) and it is generally hard to directly generalize these approaches to more comprehensive nonlinear frameworks and benefit from the computational power that the deep learning model brought to us. One class of these methods relies on side measurements such as single-cell chromatin accessibility or transcription factor (TF) binding motifs (Kamimoto et al., 2023). However, these measurements often require more complicated experimental designs and could also introduce additional noise as these data could come from different experiments.

To address the above problems, we present MCNET, a deep generative model that can jointly embed the gene expression data and simultaneously construct a GRN that reflects the inner structure of gene interactions in single cells. To implement such an idea, we heavily based our work on the work of [29], which generalized a popular approach, called the structural equation model (SEM), that infers the causality using a linear model, and implemented the exact technique that was desired for our project. (Shu et al., 2021) hypothesised that by adding proper mathematical constraints, part of the neural network architecture could be used to predict the GRN of the scRNA-seq data. A previous study by (Lin et al., 2017) showed that more accurate cell representations
could be achieved by guiding the neural network architecture with a GRN structure derived from the literature and databases. In this Article, we show that the neural network architecture can reflect GRN structure by properly designing the neural network layer with a reliance on multi omic data. Integration of multi-omic datasets in a Structural Equation Modelling neural network was based on the work of (Picard et al., 2021). The neural network architecture can be inferred jointly with the training of the weights of the neural network in an end-to-end manner.

We evaluate the performance of MCNET for various single-cell tasks such as GRN inference, scRNA-seq data visualization, cell-type identification and cell simulations on several benchmark datasets. We first show that MCNET is able to achieve better performance on the GRN inference task compared with the state-of-the-art algorithms on several popular benchmark datasets. We also apply MCNET to another single-cell dataset without the ground-truth GRN measured, and provide extensive evidence extracted from the single-cell DNA methylation and open chromatin data to demonstrate the accuracy and efficiency of our algorithm. Moreover, we also evaluate the quality of the single-cell representation reg-

ularized by the GRN structure. We find that MCNET can achieve comparable or better performance compared with current state-of-the-art methods on the tasks of visualization and cell-type identification on various benchmark datasets.

2. INTEGRATION OF MULTI OMIC DATA

The advent of powerful and inexpensive screening technologies (Misra et al., 2019) recently produced huge amounts of biological data that opened the way to a new era of therapeutics and personalized medicine (Ahmed, 2020) Treatment efficiency and adverse effects can differ vastly between individuals due to differences in age, sex, genetics and environmental factors (e.g., anthropometric and metabolic statuses; dietary and lifestyle habits (Burney & Lakhtakia, 2017; Jaccard et al., 2017) The aim of precision medicine is thus to design the most appropriate intervention based on the biological information of each individual (Tebani et al., 2016). Clinical information and omics data can be directly retrieved from databases or collected with screening technologies for disease (Menyhárt & Győrffy, 2021), class prediction (Hasin et al., 2017), biomarkers discovery (Sun & Hu, 2016), disease subtyping (Menyhárt & Győrffy, 2021), improved system biology knowledge (Dahal et al., 2020), drug repurposing and so on. Each type of omics data is specific to a single "layer" of biological information such as genomics, epigenomics, transcriptomics, proteomics, metabolomics, and provides a complementary medical perspective of a biological system or an individual (Misra et al., 2019). In the past, single-omics studies were done in hope of discovering the causes of pathologies and helping select an appropriate treatment. We now realize that such approaches are overly simplistic. Most diseases affect complex molecular pathways where different biological layers interact with each other. Hence the need for multi-omics studies that can encompass several layers at once and draw a more complete picture of a given phenotype (Tian et al., 2014) With multiple omics, faint patterns in gene expression data can be reinforced with epigenomics (Zarayeneh et al., 2017) for example. Complementary information can be exploited to better explain classification results (Rappoport et al., 2020), improve prediction performances (Sharifi-Noghabi et al., 2019), (Tini et al., 2017) or understand complex molecular pathways (Akhmedov et al., 2017) that would be out of grasp for single-omics studies. However, multi-omics studies include data that differ in type, scale and distribution, with often thousands of variables and only few samples. Additionally, biological datasets are complex, noisy, with potential errors due to measurement mistakes or unique biological deviations. Discovering pertinent information and integrating the omics into a meaningful model is therefore difficult and a great number of methods and strategies have been developed in recent years to tackle this challenge (Menyhárt & Győrffy, 2021), (Higdon et al., 2015). If the integration is not done correctly, adding more omics might not result in a significant increase of performance, but will increase the complexity of the problem along with computational time.

2.1. Main integration strategies

From multiple omics datasets, each having the same rows representing samples (patients, cells) and different columns representing biological variables grouped by omics (gene expression, copy number variation, miRNA expression, etc.), different goals could be achieved such as sample classification, disease subtyping, biomarker discovery, etc. Machine learning (ML) models are commonly used to analyze complex data, but the integration of multiple noisy and highly dimensional datasets is not straightforward. Hence, multiple integration strategies have been developed, each one of them having pros and cons. Assuming each dataset has been pre-processed according to its omics data, the datasets could simply be assembled with sample wise concatenation and the resulting matrix used as input to ML models. But in practice, most ML models will struggle to learn on such a complex dataset, particularly if the number of samples is low. Other strategies rely on transforming or mapping the datasets to reduce their complexity, either independently or jointly. An opposite strategy can also be adopted , which does not combine data and analyzes each omics dataset separately. The prediction of each model is assembled afterward for a final decision. Finally, the hierarchical strategy integrates the omics datasets by taking into account the known regulatory relationships between omics as presented by the central dogma of molecular biology (CRICK, 1970).

2.2. Hierarchical integration

A challenge in system biology is to understand the modular organization structured at the molecular level. A new trend is to incorporate these regulatory effects in the integration strategy to better reflect the nature of multidimensional data. Hierarchical strategy bases the multi-omics integration on the inclusion of the prior knowledge of regulatory relationships between the different layers. For example, a strategy for genotype-phenotype integration based on existing knowledge of cellular subsystems could follow this logic: genotypic variations in nucleotides can give rise to change in gene expression or functional changes in proteins which in turn could ultimately affect the phenotype. Therefore, hierarchical integration strategies often use external information from interaction databases and scientific literature. Moreover, because omics are organized in sequential fashion, the challenges of multi-omics integration are not exacerbated and can be dealt with separately for each dataset. Some methods for supervised hierarchical integration include Bayesian analysis of genomics data (iBAG) (Wang et al., 2013), linear regulatory modules (LRMs) (Zhu et al., 2016), and Assisted Robust Marker Identification (ARMI) (Chai et al., 2017), and Robust Network (Wu et al., 2018). Hierarchical integration meth-

ods are often designed to study specific regulatory relationships. For example, iBAG has been developed to investigate associations between epigenetic and gene expression regulation. The framework uses hierarchical modeling to combine the data from methylation and gene expression to study the associations with patient survival. Robust Network has developed an approach for modeling the gene expression (GE) and copy number variation (CNV) regulation that describe the dominant cis-acting CNV effects compared to trans-acting CNVs. This approach could be extended to other regulation relationships such as gene expression by methylation and microRNAs. Additionally, hierarchical integration can be used to infer gene regulatory networks (GRN) from multi-omics datasets.

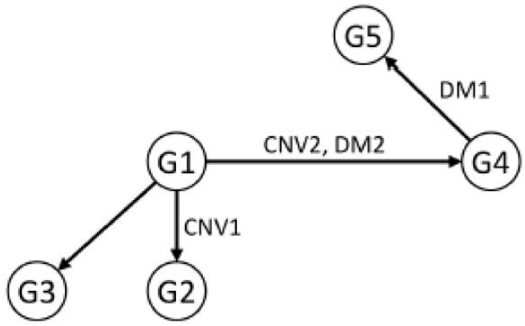


Figure 1. A simple integrative gene regulatory network. The interaction effects of copy number variation (CNV) and DNA methylation (DM) are incorporated in the gene regulatory network model

3. METHODS

3.1. Notations

We describe the notations used throughout this paper. Let G denote a matrix of gene expression, $G \in \mathbb{R}^{N\times P}$, where N is the number of samples and P is the number of genes in microarray data. The vector of gene expression at the ith gene is denoted as $g_i \in \mathbb{R}^N$, and G_{-i} represents the matrix that contains gene expressions other than gene i, $G_{-i} \in \mathbb{R}^{N\times(P-1)}$. The matrices of CNV and DNA methylation data are denoted as $C \in \mathbb{R}^{N\times V}$ and $D \in \mathbb{R}^{N\times M}$, respectively. We suppose that CNVs or DNA methylations can be annotated to nearby genes (upstream or downstream of the gene) in the same chromosome. The gene annotations of the CNV and DNA methylation to gene i are represented as the matrices $C(i) \in \mathbb{R}^{N\times V_i}$ and $D_i \in \mathbb{R}^{N\times M_i}$, respectively, where V_i and M_i are the numbers of CNVs and DNA methylations that are annotated to gene i.

The regulatory relationships between genes are represented by an adjacency matrix of gene expression $B \in \mathbb{R}^{P\times P}$, and integrative interactions of multi-omics data other than gene expression are expressed by their own biadjacency matrices. In MCNET, the interactions of CNVs and DNA methylations to genes are described as $BC \in \mathbb{R}^{V\times P}$ and $BD \in \mathbb{R}^{M\times P}$, respectively. We assume that there is no self-regulation in the gene regulatory network, i.e., $B_{ii} = 0$, where $i = \{1, \ldots, P\}$.

3.2. Integrative gene regulatory network inference

We propose an Integrative Gene Regulatory Network inference (MCNET) method that infers a gene regulatory network from multi-omics data. The current state-of-the-art methods for integrative gene regulatory network inference using multi-omics data, such as SGRN (Cai et al., 2013) and DCGRN (D.-C. Kim et al., 2014a), consider all the types of data as nodes in networks. In other words, nodes can indicate genes, CNVs, or DNA methylations. In contrast, MCNET constructs homogeneous gene regulatory networks where nodes represent only genes, which consequently makes it possible to apply most graph algorithms for further analysis. Figure 1 shows a simple integrative gene regulatory network, where a gene (G1) regulates another gene (G4) with biological processes of a CNV (CNV2) and a DNA methylation (DM2). The proposed method, MCNET, represents integrative gene regulatory networks with multi-layered adjacency matrices of the multi-omics data. It constructs the adjacency matrix of gene expression and the biadjacency matrices of CNV and DNA methylation. The adjacency matrix of gene expression defines the basic structure of the transcriptional biological networks, and the biadjacency matrices of CNV and DNA methylation describe their integrative interactions on the gene regulations.
For formulating the integration of the heterogeneous data into a standardised format, MCNET takes into account the interaction effects of CNVs and DNA methylations with genes. The integrative interactions between a gene i and its nearby CNVs and DNA methylations can be described by Fisher's interaction model:

$$\mathbf{g}_i \otimes \mathbf{C}_i, \quad \mathbf{g}_i \otimes \mathbf{D}_i \tag{1}$$

where $\otimes$ is an element-by-element multiplication. It explains different gene expression levels on the variations of CNVs or DNA methylations. Thus, the expression of gene i can be represented by a sparse linear model by incorporating not only other genes but also interaction effects of its nearby CNVs and DNA methylations. The gene expression (g_i) for gene i is formulated by:

$$\begin{aligned} \mathbf{g}_i = \mathbf{G}_{-i}\mathbf{b}_{ig} + (\mathbf{C}_i \odot \mathbf{g}_i)\mathbf{b}_{ic} \\ + (\mathbf{D}_i \odot \mathbf{g}_i)\mathbf{b}_{id} + \varepsilon_i, \\ \text{subject to} \quad |\mathbf{b}_{ig}| \le \mathbf{C}_g, \quad |\mathbf{b}_{ic}| \le \mathbf{C}_c, \quad |\mathbf{b}_{id}| \le \mathbf{C}_d, \end{aligned} \tag{2}$$

where b_{gi}, b_{ci}, and b_{di} are the coefficients of gene expressions other than gene i, CNVs, and DNA methylations of gene i, respectively. $|\cdot|$ is the L-1 norm, and the residual is denoted as ε_i. The adjacency matrix B of the gene regulatory network is comprised of b_{gi} $(1 \le i \le P)$ in (2), i.e.,

$B = \{b_{g1}, \ldots, b_{gP}\}^{\perp}$.

The biadjacency matrices of CNVs BC and DNA methylations BD are also constructed by b_{ci} and b_{di}.

The integrative gene regulatory network can be inferred by optimizing the parameters of (2). The learning function $F(b_{gi}, b_{ci}, b_{di})$ for the optimal parameters is obtained by using least squares with the sparse setting:

$$\begin{aligned}\text{argmin } F(\mathbf{b}_i^g, \mathbf{b}_i^c, \mathbf{b}_i^d) = &\left\|\mathbf{g}_i - \left(\mathbf{G}_{-i}\mathbf{b}_i^g + \left(\mathbf{C}_{\{i\}} \otimes \mathbf{g}_i\right)\mathbf{b}_i^c\right.\right. \\ &\left.\left. + \left(\mathbf{D}(i) \otimes \mathbf{g}_i\right)\mathbf{b}_i^d\right)\right\|_2 \\ &+ \lambda_g\|\mathbf{b}_i^d\| + \lambda_c\|\mathbf{b}_i^c\| + \lambda_d\|\mathbf{b}_i^d\|,\end{aligned} \tag{3}$$

where λ_g, λ_c, and λ_d are hyper-parameters for sparsity regularization, and $|\cdot|_2$ is the L-2 norm. The optimization function can be considered as the following LASSO problem:

$$\arg\min |\mathbf{g}_i - \mathbf{X}\mathbf{b}_i|_2 + \lambda |\mathbf{b}_i| \tag{4}$$

where X is the augmented matrix, $X = \{G_{-i}, C(i) \otimes g_i, D_i \otimes g_i\}$. However, the number of genes in $(P-1)$ is much larger than the number of CNVs (V_i) and DNA methylations (M_i) associated with gene i. For instance, there are only a couple of CNVs (C_i) or DNA methylations (D_i) for a gene in the psychiatric disorder data that we used for the experiment in the paper, whereas the number of genes in G_{-i} is in the hundreds even after pre-processing. Thus, the solution of LASSO may tend to ignore most CNVs and DNA methylations despite their importance. Therefore, we solve the optimization problem in a stepwise manner. First, we identify significant genes that interact with gene i from G_{-i} by LASSO:

$$\arg\min_{\boldsymbol{g}_i} \left\|\boldsymbol{g}_i - \boldsymbol{G}_{-i}^{\boldsymbol{b}_{ig}}\right\|^2 + \lambda \left\|\boldsymbol{b}_i^g\right\|. \tag{5}$$

The matrix of G'_{-i} is constructed with the genes with non-zero coefficients. Secondly, we compute p-values of the variables in the following linear regression:

$$\boldsymbol{g}_i = \boldsymbol{G}_{-i}^{\boldsymbol{b}'^g_{ig}} + (\boldsymbol{C}_i \otimes \boldsymbol{g}_i)^{\boldsymbol{b}_i^c} + (\boldsymbol{D}_i \otimes \boldsymbol{g}_i)^{\boldsymbol{b}_i^d} + \boldsymbol{\varepsilon}_i. \tag{6}$$

The coefficients of the genes, CNVs, and DNA methylations with p-values ≥ 0.05 are set to zeros. Then, the coefficients for genes are assigned to the adjacency matrix, and b_{ci} and b_{di} are assigned to the biadjacency matrices of BC and BD respectively. The procedure is described in Algorithm 1.

Algorithm 1 Algorithm 1

1: **For** $i \in \{1, \ldots, P\}$ **do**
2: $b_{gi} = \text{LASSO}(G - i, g_i)$
3: *Compute the linear regression of (6)*
4:

$$b'^g_{ij} = \begin{cases} b'^g_{ij} & \text{if } b'^g_{ij} \text{ is non-zero and p-value}(b'^g_{ij}) < 0.05 \\ 0 & \text{otherwise} \end{cases}$$

5:

$$b^c_{ij} = \begin{cases} b^c_{ij} & \text{if } if\ b^c_{ij}\ is \text{ non-zero and p-value}(b^c_{ij}) < 0.05 \\ 0 & \text{otherwise} \end{cases}$$

6:

$$b^d_{ij} = \begin{cases} b^d_{ij} & \text{if } if\ b^d_{ij}\ is \text{ non-zero and p-value}(b^d_{ij}) < 0.05 \\ 0 & \text{otherwise} \end{cases}$$

7: Construct B, B_C, and B_D
8: ***end for***

4. SIMULATION STUDIES

We conducted intensive simulation experiments to evaluate our proposed method and compare the performance with existing methods. Due to only few available well-known true models of biological networks, the assessment of gene regulatory network inference in complex organisms such as human is challenging. Thus, the performance was indirectly evaluated with simulation data that implements integrative biological networks where the true model is given.

We generated the simulation data under the assumption that we hypothesised for the integrative gene regulatory networks. In the simulation studies, we aim to (1) verify that our proposed method produces robust performance to identify the true models of gene regulatory networks from multi-omics data, and (2) to compare the performance with current state-of-the-art methods on the given hypothesis. We carried out the following three experiments with the simulation data: (1) Receiver Operating Characteristic (ROC) curve, (2) sensitivity, and (3) false discovery rate.

4.1. Simulation settings

In the integrative gene regulatory network model, gene expression can be represented by two components: (1) gene expression regulated by other genes (G_g) and (2) interactions of CNVs and DNA methylations (G_i), as shown in (2):

$$G = G_g + G_i \tag{7}$$

where

$$\mathbf{G}_g = \mathbf{G}_{-i}\mathbf{b}_{ig}, \quad \mathbf{G}_i = (\mathbf{C}_{\{i\}} \otimes \mathbf{g}_i)\mathbf{b}_{ic} + (\mathbf{D}_{\{i\}} \otimes \mathbf{g}_i)\mathbf{b}_{id}.$$

First, $\mathbf{G}_g$ was generated by the given adjacency matrix $\mathbf{Z}$:

$$\mathbf{G}_g = \mathbf{E}(I - \mathbf{Z})^{-1} \tag{8}$$

Where $\mathbf{I} \in \mathbb{R}^{N \times P}$ is an identity matrix, and $\mathbf{E} \in \mathbb{R}^{N \times P}$ is a matrix with normally distributed random values for noise, $\mathbf{E} \sim \mathcal{N}(0, 0.01)$. The adjacency matrix $\mathbf{Z}$ is a sparse acyclic graph without self-loop.

The CNV data ($\mathbf{C} \in \mathbb{R}^{N \times P}$) was implemented by taking the values $\{0, 1, 2, 3, 4\}$ with the corresponding probabilities $\{0.01, 0.02, 0.4, 0.2, 0.1\}$. The given probabilities were directly acquired from CNV of human brain data. The DNA methylation ($\mathbf{D} \in \mathbb{R}^{N \times M}$) was randomly obtained by the uniform distribution on the interval $[0, 1]$. In practice, CNVs and DNA methylations were annotated to nearby genes by using their loci and gene regions. We designated the associations by sparse Boolean mapping matrices $\mathbf{W} \in \mathbb{R}^{V \times P}$ and $\mathbf{F} \in \mathbb{R}^{M \times P}$ for CNVs and DNA methylations, where only a couple of CNVs and DNA methylations can be annotated to a gene. In this simulation data, we assume that all of the CNVs and DNA methylations nearby a gene significantly regulate the gene expression.

The gene expression regulated by the interactions of CNVs and DNA methylations was generated by:

$$\mathbf{G}_i = \mathbf{CW} \otimes \mathbf{G} + \mathbf{DF} \otimes \mathbf{G}, \quad (9) \tag{9}$$

The gene expression controls the gene expression levels of other genes with the interaction effects of multi-omics data in gene regulatory networks. Therefore, we repeated Equation (8) and Equation (9) until $\mathbf{G}$ converges. Note that $\mathbf{Z}$, $\mathbf{W}$, and $\mathbf{F}$ are the (bi)adjacency matrices of ground truth in the simulation studies. The algorithm is described in Algorithm 2.

Algorithm 2 Algorithm 2

1: $G = E(I - Z)^{-1}$
2. ***do***
3. $\quad \mathbf{G} = (\mathbf{E} + \mathbf{CW} \otimes \mathbf{G} + \mathbf{DF} \otimes \mathbf{G})(\mathbf{I} - \mathbf{Z})^{-1}$
4. ***while G** converges*

We considered a LASSO-based GRN method (GRN) as baseline and DCGRN (D.-C. Kim et al., 2014b) which is an integrative gene regulatory network inference method that uses multi-omics data. GRN infers the gene regulatory relationship on gene i with LASSO regularisation:

$$g_i = G^{-i} b_{gi} + \epsilon_i, \quad \text{subject to} \quad |b_{gi}| < C_g \tag{10}$$

GRN identifies significant gene regulations by LASSO solution, but it considers only gene expression data for the network inference. In contrast, DCGRN incorporates multi-omics data of CNVs and DNA methylations in the model:

$$g_i = G^{-i} b_{gi} + C(i) b_{ic} + D(i) b_{id} + \epsilon_i \quad \text{subject to } |b_{gi}| \le C_g, \quad |b_{ic}| \le C_c, \quad \text{and} \quad |b_{id}| \le C_d. \tag{11}$$

4.2. Experimental results with simulation data

First, we evaluated the performance by computing the area under the receiver operating characteristic curve (AUROC). The confusion matrix of true positive (TP), false positive (FP), true negative (TN), and false negative (FN) is defined as:

- TP: correctly identified the positive gene regulations as non-zero coefficients,
- FP: incorrectly identified the positive gene regulations as zero coefficients,
- TN: correctly identified the negative gene regulations as zero coefficients,
- FN: incorrectly identified the negative gene regulations as non-zero coefficients.

The non-zero coefficients of b_{gi}, b_{di}, and b_{ci} were considered as positives, while the coefficients of zero were negatives. The confusion matrices for gene regulations and integrative interactions of CNVs and DNA methylations were separately computed.

The ROC curves were traced over different thresholds to examine the trade-off between True Positive Rate (TPR = TP/(TP+FN)) and False Positive Rate (FPR = FP/(FP+TN)). The hyper-parameters (λ_g, λ_c, and λ_d) in (3) determine the sparsity of significant components with non-zero coefficients in the multi-omics data. Note that all of the coefficients are non-zero when the parameter is zero, while all coefficient values become zero when an infinite value is given for the parameter. We considered the sparsity step ($1 \le \theta \le P + V + M$) that determines the hyper-parameters in the LASSO solution. In this simulation study for the ROC curves, only the coefficient values were considered to determine the positive interactions, where p-values were not computed.

GRN computes only a confusion matrix for gene regulations, while DCGRN and MCNET have confusion matrices for CNVs and DNA methylations as well as gene expression. Therefore, overall ROC curves were considered, where only the confusion matrix of gene regulation was reflected on GRN, while the three confusion matrices were combined to compute ROC curves in DCGRN and MCNET. The overall ROC curves are illustrated in Figure 2, and AUROC is shown in Table 1. The experimental result of the overall AUROC supports that MCNET (0.938) provides better performance than GRN (0.895) and DCGRN (0.843).

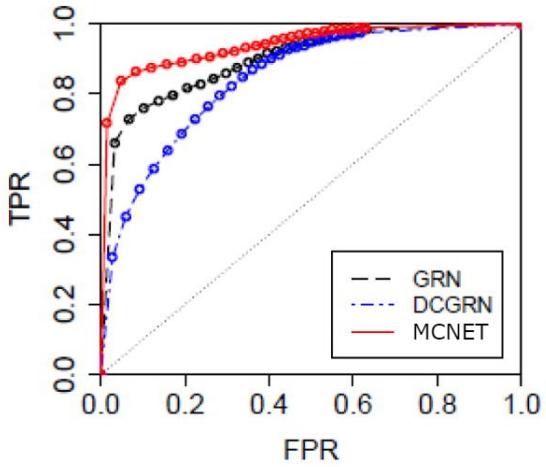


Figure 2. Overall ROC curves

Table 1. AUROC with simulation data

Methods	GRN	DCGRN	MCNET
AUROC	0.895	0.843	0.938

TPRs on interactions of the CNVs and DNA methylations were measured for DCGRN and MCNET. Since the simulation data does not include negatives on CNVs and DNA methylations, we examined how well the methods identify the true positives. The TPRs are shown in Figure 3, where MCNET outperforms DCGRN in identifying true integrative interactions of CNVs and DNA methylations.

Secondly, we measured the overall sensitivity which is the probability of identifying the true positives. In this simulation study, the hyper-parameters were optimised by 10-fold cross-validation. The multi-omics elements with non-zero coefficients and whose p-values are less than 0.05 are considered as positives. The overall sensitivity is depicted in Figure 4. MCNET produced the best sensitivity (0.300 ± 0.034), while GRN and DCGRN showed 0.199 ± 0.030 and 0.269 ± 0.035 respectively. The sensitivities of CNVs and DNA methylations on MCNET and DCGRN are shown in Figure 5. The sensitivities for MCNET and DCGRN were 0.102 ± 0.035 and 0.054 ± 0.030, respectively.

Lastly, we conducted the simulation study for False Discovery Rate (FDR). In this study, we generated simulation data that had no gene-gene regulation in the biological network. All positive predictions inferred by the methods were false positives, as the true adjacency matrix consisted entirely of zeros. FDR was computed as FP/(TP + FP). The FDRs of GRN, DCGRN, and MCNET, which were observed in Figure 6, were all less than 0.02. Specifically, the FDRs were 0.019 ± 0.003, 0.019 ± 0.003, and 0.019 ± 0.003 for GRN, DCGRN, and MCNET, respectively. These results indicate that MCNET had a chance of misidentifying interactions of

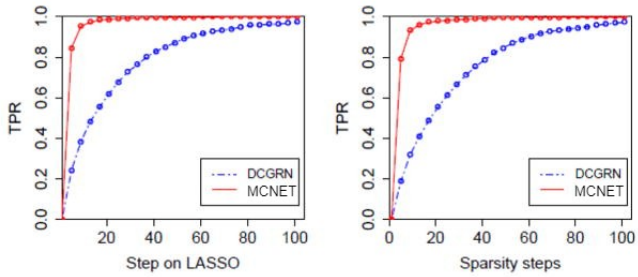


Figure 3. TPRs for interaction effects of CNVs and DNA methylations

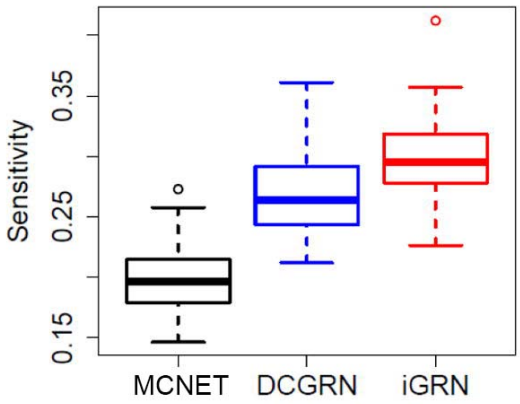


Figure 4. Sensitivity

less than 2

5. CONCLUSION

The promise of deep learning, based on its success in other fields (Krizhevsky et al., 2017), is now also being seen across many different areas of biological network analysis. The methods we reviewed reported to consistently match or beat previous state-of-the-art methods using classical machine learning algorithms, providing evidence of one of deep learning's core advantages: its strong empirical classification performance. Another advantage of deep learning is its ability to effectively deal with large datasets , which can be challenging for classical machine learning methods (Zhou et al., 2017) Although the training process of deep learning models with huge amounts of data is a non-trivial task, the advances in parallel and distributed computing have made training these large deep learning models possible (Dean et al., 2012; LeCun et al., 2015). The large number of matrix multiplications, high memory requirements and easy parallelizability of neural networks have been particularly well served by the recent breakthroughs in GPU computing (**?**). Finally, given that deep learning is a learning approach based on a hierarchy of non-linear functions, it is capable of detecting patterns in the raw data without explicit feature engineering. While it is not the only method that can handle non-linear re-

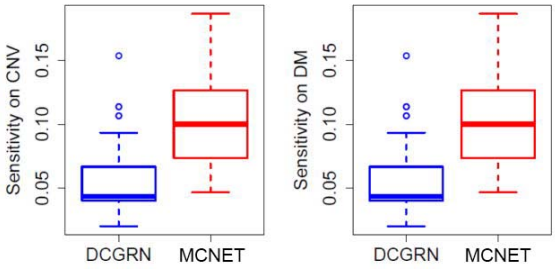


Figure 5. Sensitivity on copy number variations and DNA methylations

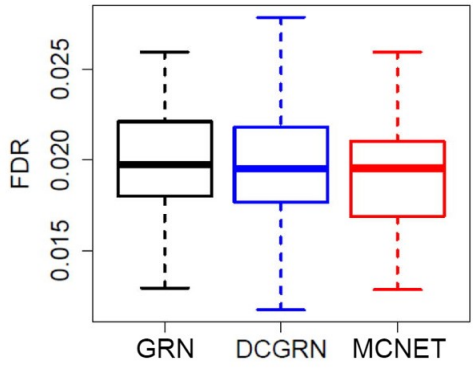


Figure 6. False discovery rate

lationships, the composition of many simple, non-linear layers makes it particularly adept at learning patterns at different layers of abstraction (LeCun et al., 2015), enabling more complex patterns to be detected. While deep learning methods are very promising, there are limitations and many open questions to be solved. One of the main problems with deep learning is its lack of interpretability. While there has been some recent progress in this area (Ching et al., 2018; Li et al., 2019), the black box nature of deep learning algorithms remains a key challenge, particularly in bioinformatics, where one is interested in understanding the mechanisms underlying the biological processes (Miotto et al., 2018; Zampieri et al., 2019). Additionally, interpretability is critical in the context of models that guide medical decisions, where doctors and patients are often unlikely to trust the output of a deep learning model without sufficient understanding of the prediction process (Ching et al., 2018).

Multi-omics data can be used in modeling gene regulatory networks. The recent rapid advances of high-throughput omics technologies have triggered the integrative multi-omics study for the in-depth understanding of the complex biological processes. However, only a few studies have considered the multi-omics data in gene regulatory network inference.

In this paper, we proposed an integrative gene regulatory network inference method, where multi-omics data and their interaction effects are integrated in the mathematical graph model. Our proposed method, MCNET, can infer gene regulatory networks from multi-omics data of CNVs and DNA methylations as well as gene expression data, and produce the homogeneous network where nodes are only genes. It enables one to analyse the gene regulatory network with most network analysis and visualisation tools efficiently. The inference capability of MCNET was assessed by the intensive experiments with simulation data. MCNET was applied to human brain data of psychiatric disorders, and the biological network of psychiatric disorders was analysed.

REFERENCES

Ahmed, Z. (2020, October). Practicing precision medicine with intelligently integrative clinical and multi-omics data analysis. *Human Genomics*, *14*(1). doi: 10.1186/s40246-020-00287-z

Akhmedov, M., Arribas, A., Montemanni, R., Bertoni, F., & Kwee, I. (2017). Omicsnet: Integration of multi-omics data using path analysis in multilayer networks. *bioRxiv*, 238766.

Blencowe, B. J., Ahmad, S., & Lee, L. J. (2009, June). Current-generation high-throughput sequencing: deepening insights into mammalian transcriptomes. *Genes & Development*, *23*(12), 1379–1386. doi: 10.1101/gad.1788009

Burney, I. A., & Lakhtakia, R. (2017). Precision medicine: Where have we reached and where are we headed? *Sultan Qaboos University Medical Journal*, *17*(3), e255.

Cai, X., Bazerque, J. A., & Giannakis, G. B. (2013, May). Inference of gene regulatory networks with sparse structural equation models exploiting genetic perturbations. *PLoS Computational Biology*, *9*(5), e1003068. doi: 10.1371/journal.pcbi.1003068

Cases, I., & de Lorenzo, V. (2005, February). Promoters in the environment: transcriptional regulation in its natural context. *Nature Reviews Microbiology*, *3*(2), 105–118. doi: 10.1038/nrmicro1084

Chai, H., Shi, X., Zhang, Q., Zhao, Q., Huang, Y., & Ma, S. (2017, September). Analysis of cancer gene expression data with an assisted robust marker identification approach. *Genetic Epidemiology*, *41*(8), 779–789. doi: 10.1002/gepi.22066

Chan, T. E., Stumpf, M. P., & Babtie, A. C. (2017). Gene regulatory network inference from single-cell data using multivariate information measures. *Cell systems*, *5*(3), 251–267.

Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., ... others (2018). Opportunities and obstacles for deep learning in biology and

medicine. *Journal of The Royal Society Interface*, *15*(141), 20170387.

CRICK, F. (1970, August). Central dogma of molecular biology. *Nature*, *227*(5258), 561–563. doi: 10.1038/227561a0

Dahal, S., Yurkovich, J. T., Xu, H., Palsson, B. O., & Yang, L. (2020). Synthesizing systems biology knowledge from omics using genome-scale models. *Proteomics*, *20*(17-18), 1900282.

Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Mao, M., ... others (2012). Large scale distributed deep networks. *Advances in neural information processing systems*, *25*.

García de la Nava, J., van Hijum, S., & Trelles, O. (2004). Saturation and quantization reduction in microarray experiments using two scans at different sensitivities. *Statistical Applications in Genetics and Molecular Biology*, *3*(1).

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Hasin, Y., Seldin, M., & Lusis, A. (2017, May). Multi-omics approaches to disease. *Genome Biology*, *18*(1). doi: 10.1186/s13059-017-1215-1

Hempel, S., Koseska, A., Nikoloski, Z., & Kurths, J. (2011, July). Unraveling gene regulatory networks from time-resolved gene expression data – a measures comparison study. *BMC Bioinformatics*, *12*(1). doi: 10.1186/1471-2105-12-292

Higdon, R., Earl, R. K., Stanberry, L., Hudac, C. M., Montague, E., Stewart, E., ... Kolker, E. (2015, April). The promise of multi-omics and clinical data integration to identify and target personalized healthcare approaches in autism spectrum disorders. *OMICS: A Journal of Integrative Biology*, *19*(4), 197–208. doi: 10.1089/omi.2015.0020

Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., & Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. *PloS one*, *5*(9), e12776.

Jaccard, E., Cornuz, J., Waeber, G., & Guessous, I. (2017, August). Evidence-based precision medicine is needed to move toward general internal precision medicine. *Journal of General Internal Medicine*, *33*(1), 11–12. doi: 10.1007/s11606-017-4149-0

Jacob, F., & Monod, J. (1961, June). Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, *3*(3), 318–356. Retrieved from https://doi.org/10.1016/s0022-2836(61)80072-7 doi: 10.1016/s0022-2836(61)80072-7

Kamimoto, K., Stringa, B., Hoffmann, C. M., Jindal, K., Solnica-Krezel, L., & Morris, S. A. (2023). Dissecting cell identity via network inference and in silico gene perturbation. *Nature*, *614*(7949), 742–751.

Karlebach, G., & Shamir, R. (2008, September). Modelling and analysis of gene regulatory networks. *Nature Reviews Molecular Cell Biology*, *9*(10), 770–780. Retrieved from https://doi.org/10.1038/nrm2503 doi: 10.1038/nrm2503

Kim, D.-C., Kang, M., Zhang, B., Wu, X., Liu, C., & Gao, J. (2014a). Integration of dna methylation, copy number variation, and gene expression for gene regulatory network inference and application to psychiatric disorders. In *2014 ieee international conference on bioinformatics and bioengineering* (pp. 238–242).

Kim, D.-C., Kang, M., Zhang, B., Wu, X., Liu, C., & Gao, J. (2014b). Integration of dna methylation, copy number variation, and gene expression for gene regulatory network inference and application to psychiatric disorders. In *2014 ieee international conference on bioinformatics and bioengineering* (pp. 238–242).

Kim, S. (2015). ppcor: an r package for a fast calculation to semi-partial correlation coefficients. *Communications for statistical applications and methods*, *22*(6), 665.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90.

LeCun, Y. (1985). Une procedure d'apprentissage ponr reseau a seuil asymetrique. *Proceedings of Cognitiva 85*, 599–604.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436–444.

Li, Y., Huang, C., Ding, L., Li, Z., Pan, Y., & Gao, X. (2019). Deep learning in bioinformatics: Introduction, application, and perspective in the big data era. *Methods*, *166*, 4–21.

Lin, C., Jain, S., Kim, H., & Bar-Joseph, Z. (2017). Using neural networks for reducing the dimensions of single-cell rna-seq data. *Nucleic acids research*, *45*(17), e156–e156.

López-Barneo, J., Pardal, R., & Ortega-Sáenz, P. (2001). Cellular mechanism of oxygen sensing. *Annual Review of Physiology*, *63*(1), 259-287. (PMID: 11181957) doi: 10.1146/annurev.physiol.63.1.259

Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., ... Stolovitzky, G. (2012, July). Wisdom of crowds for robust gene network inference. *Nature Methods*, *9*(8), 796–804. doi: 10.1038/nmeth.2016

Matsumoto, H., Kiryu, H., Furusawa, C., Ko, M. S. H., Ko, S. B. H., Gouda, N., ... Nikaido, I. (2017, April). SCODE: an efficient regulatory network inference algorithm from single-cell RNA-seq during differentiation. *Bioinformatics*, *33*(15), 2314–2321. doi: 10.1093/bioinformatics/btx194

Menyhárt, O., & Győrffy, B. (2021). Multi-omics approaches in cancer research with applications in tumor subtyping, prognosis, and diagnosis. *Computational and*

Structural Biotechnology Journal, *19*, 949–960. doi: 10.1016/j.csbj.2021.01.009

Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2018). Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, *19*(6), 1236–1246.

Misra, B. B., Langefeld, C., Olivier, M., & Cox, L. A. (2019, January). Integrated omics: tools, advances and future approaches. *Journal of Molecular Endocrinology*, *62*(1), R21–R45. doi: 10.1530/jme-18-0055

Moerman, T., Aibar Santos, S., Bravo González-Blas, C., Simm, J., Moreau, Y., Aerts, J., & Aerts, S. (2019). Grnboost2 and arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics*, *35*(12), 2159–2161.

Omony, J. (2014, January). Biological network inference: A review of methods and assessment of tools and techniques. *Annual Research & Review in Biology*, *4*(4), 577–601. doi: 10.9734/arrb/2014/5718

Papili Gao, N., Ud-Dean, S. M., Gandrillon, O., & Gunawan, R. (2018). Sincerities: inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. *Bioinformatics*, *34*(2), 258–266.

Parker, D. B. (1985). Learning logic technical report tr-47. *Center of Computational Research in Economics and Management Science, Massachusetts Institute of Technology, Cambridge, MA*.

Picard, M., Scott-Boyer, M.-P., Bodein, A., Périn, O., & Droit, A. (2021). Integration strategies of multi-omics data for machine learning analysis. *Computational and Structural Biotechnology Journal*, *19*, 3735–3746. Retrieved from https://doi.org/10.1016/j.csbj.2021.06.030 doi: 10.1016/j.csbj.2021.06.030

Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., ... Betel, D. (2013). Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biology*, *14*(9), R95. doi: 10.1186/gb-2013-14-9-r95

Rappoport, N., Safra, R., & Shamir, R. (2020, September). MONET: Multi-omic module discovery by omic selection. *PLOS Computational Biology*, *16*(9), e1008182. doi: 10.1371/journal.pcbi.1008182

Rehrauer, H., Aquino, C., Gruissem, W., Henz, S. R., Hilson, P., Laubinger, S., ... Hennig, L. (2009, December). AGRONOMICS1: A new resource for arabidopsis transcriptome profiling . *Plant Physiology*, *152*(2), 487–499. doi: 10.1104/pp.109.150185

Reuter, J., Spacek, D. V., & Snyder, M. (2015). High-throughput sequencing technologies. *Molecular Cell*, *58*(4), 586-597. doi: https://doi.org/10.1016/j.molcel.2015.05.004

Rolland, F., Baena-Gonzalez, E., & Sheen, J. (2006). Sugar sensing and signaling in plants: Conserved and novel mechanisms. *Annual Review of Plant Biology*, *57*(1), 675-709. (PMID: 16669778) doi: 10.1146/annurev.arplant.57.032905.105441

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, *323*(6088), 533–536.

Schulze, A., & Downward, J. (2001, August). Navigating gene expression using microarrays — a technology review. *Nature Cell Biology*, *3*(8), E190–E195. doi: 10.1038/35087138

Sharifi-Noghabi, H., Zolotareva, O., Collins, C. C., & Ester, M. (2019, July). MOLI: multi-omics late integration with deep neural networks for drug response prediction. *Bioinformatics*, *35*(14), i501–i509. doi: 10.1093/bioinformatics/btz318

Shu, H., Zhou, J., Lian, Q., Li, H., Zhao, D., Zeng, J., & Ma, J. (2021, July). Modeling gene regulatory networks using neural network architectures. *Nature Computational Science*, *1*(7), 491–501. Retrieved from https://doi.org/10.1038/s43588-021-00099-8 doi: 10.1038/s43588-021-00099-8

Sima, C., Hua, J., & Jung, S. (2009, September). Inference of gene regulatory networks using time-series data: A survey. *Current Genomics*, *10*(6), 416–429. doi: 10.2174/138920209789177610

Sokolov, Y., Kozma, R., Werbos, L. D., & Werbos, P. J. (2015). Complete stability analysis of a heuristic approximate dynamic programming control design. *Automatica*, *59*, 9–18.

Sun, Y. V., & Hu, Y.-J. (2016). Integrative analysis of multi-omics data for discovery and functional studies of complex human diseases. In *Advances in genetics* (pp. 147–190). Elsevier. doi: 10.1016/bs.adgen.2015.11.004

Tebani, A., Afonso, C., Marret, S., & Bekri, S. (2016). Omics-based strategies in precision medicine: toward a paradigm shift in inborn errors of metabolism investigations. *International journal of molecular sciences*, *17*(9), 1555.

Tian, X., xiao Lian, J., juan Yi, L., Ma, L., Wang, Y., Cao, H., & min Song, G. (2014, September). Current status of clinical nursing specialists and the demands of osteoporosis specialized nurses in mainland china. *International Journal of Nursing Sciences*, *1*(3), 306–313. doi: 10.1016/j.ijnss.2014.07.007

Tini, G., Marchetti, L., Priami, C., & Scott-Boyer, M.-P. (2017, December). Multi-omics integration—a comparison of unsupervised clustering methodologies. *Briefings in Bioinformatics*, *20*(4), 1269–1279. doi: 10.1093/bib/bbx167

Veal, E. A., Day, A. M., & Morgan, B. A. (2007, April). Hydrogen peroxide sensing and signaling. *Molecular Cell*, *26*(1), 1–14. doi: 10.1016/j.molcel.2007.03.016

Wang, W., Baladandayuthapani, V., Morris, J. S., Broom, B. M., Manyam, G., & Do, K.-A. (2013). ibag: integrative bayesian analysis of high-dimensional multiplatform genomics data. *Bioinformatics*, *29*(2), 149–159.

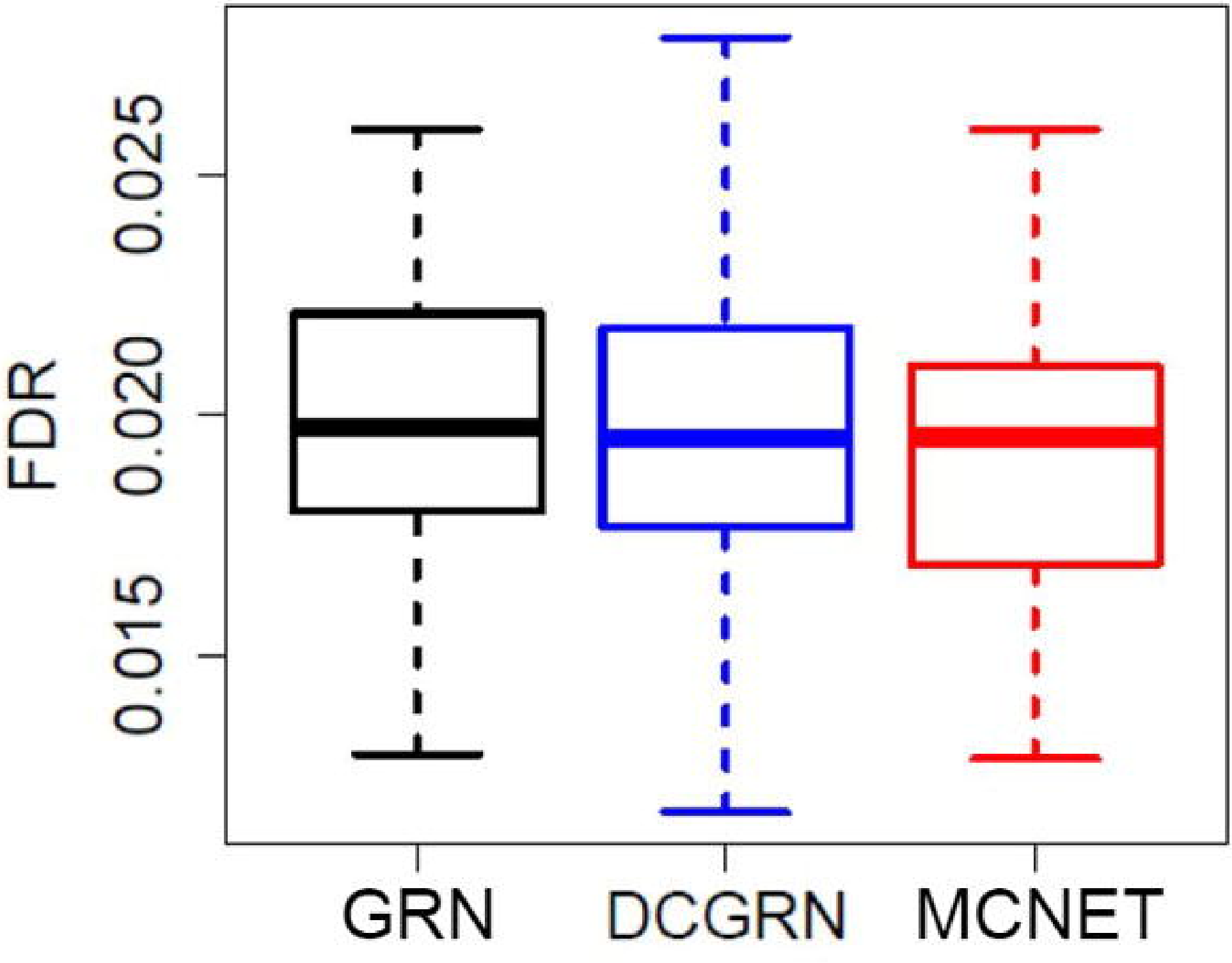
Wu, C., Zhang, Q., Jiang, Y., & Ma, S. (2018, November). Robust network-based analysis of the associations between (epi)genetic measurements. *Journal of Multivariate Analysis*, *168*, 119–130. doi: 10.1016/j.jmva.2018.06.009

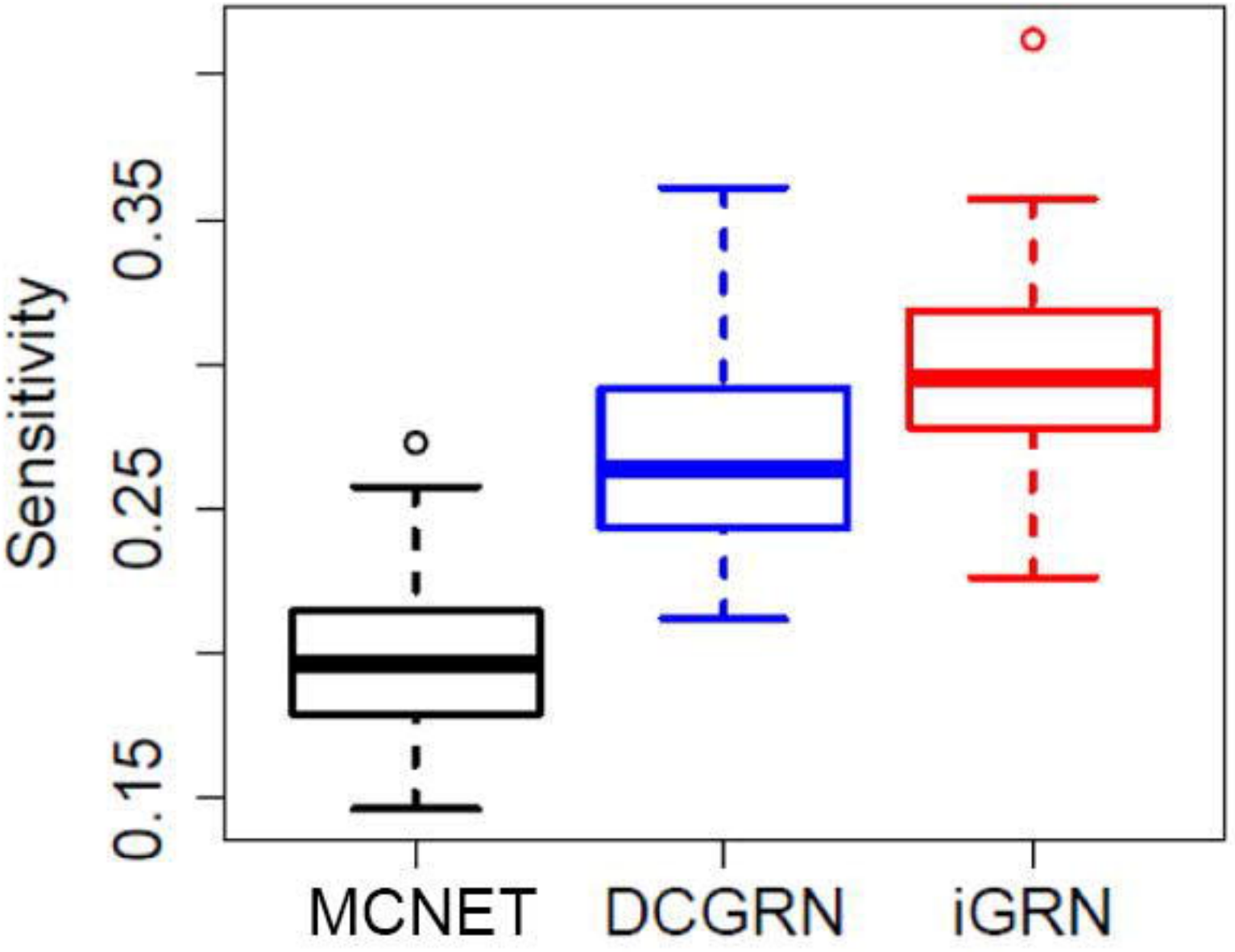
Zampieri, G., Vijayakumar, S., Yaneske, E., & Angione, C. (2019). Machine and deep learning meet genome-scale metabolic modeling. *PLoS computational biology*, *15*(7), e1007084.

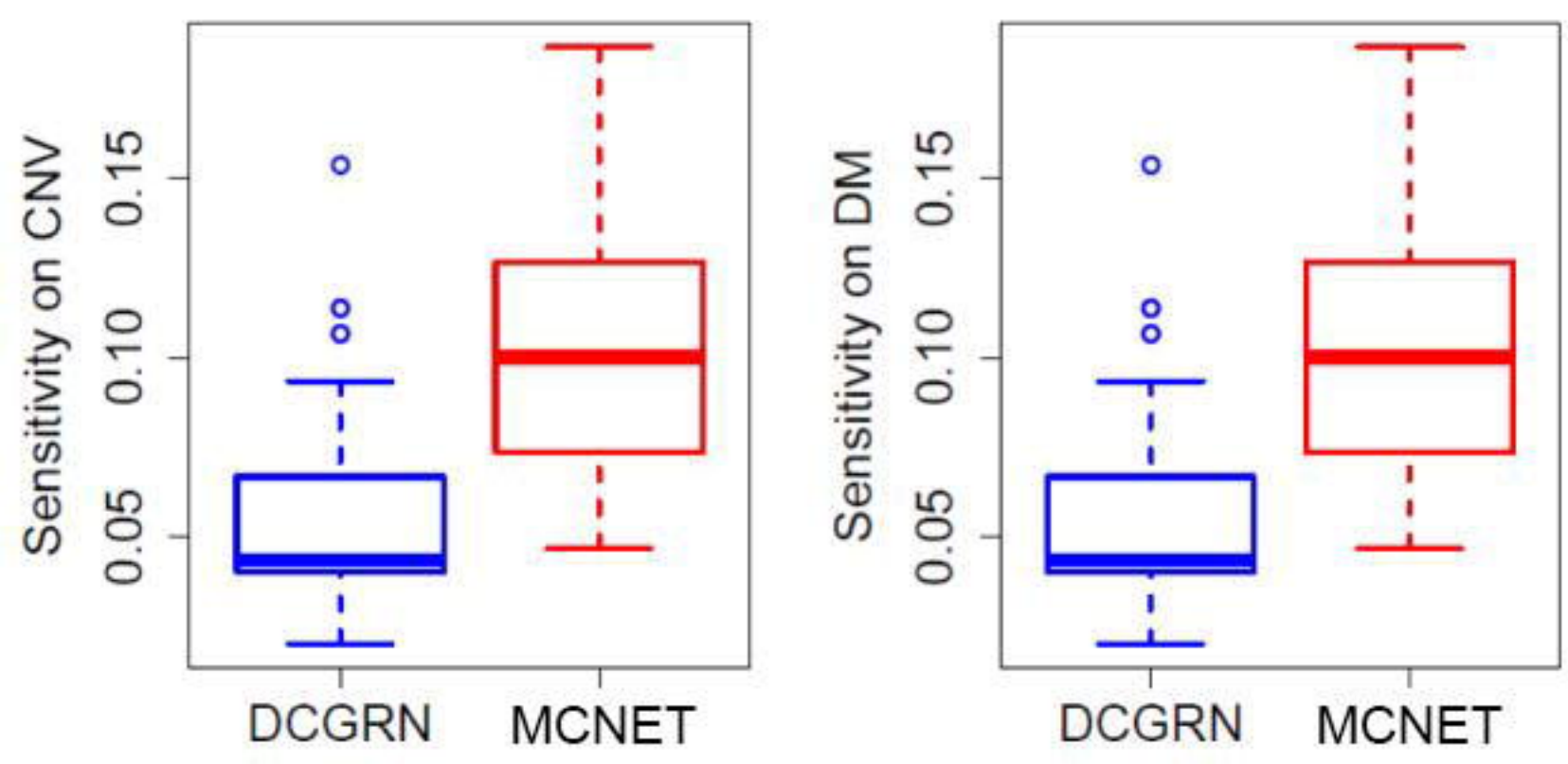
Zarayeneh, N., Ko, E., Oh, J. H., Suh, S., Liu, C., Gao, J., ... Kang, M. (2017). Integration of multi-omics data for integrative gene regulatory network inference. *International Journal of Data Mining and Bioinformatics*, *18*(3), 223. doi: 10.1504/ijdmb.2017.087178

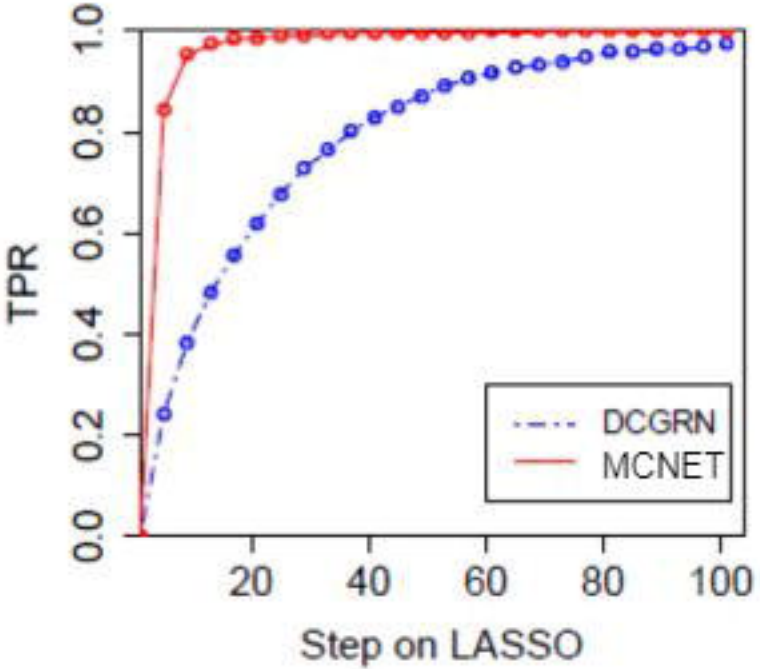
Zhou, L., Pan, S., Wang, J., & Vasilakos, A. V. (2017). Machine learning on big data: Opportunities and challenges. *Neurocomputing*, *237*, 350–361.

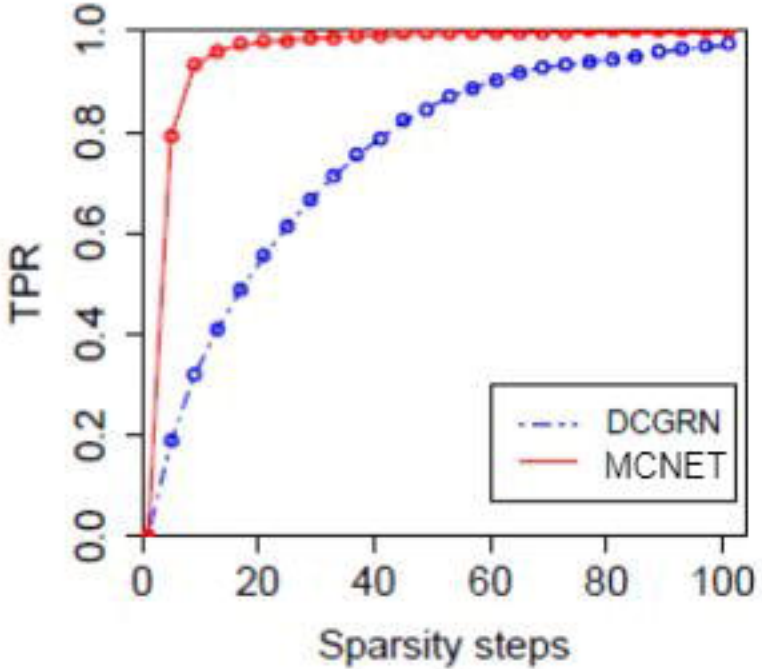
Zhu, R., Zhao, Q., Zhao, H., & Ma, S. (2016, March). Integrating multidimensional omics data for cancer outcome. *Biostatistics*, *17*(4), 605–618. doi: 10.1093/biostatistics/kxw010

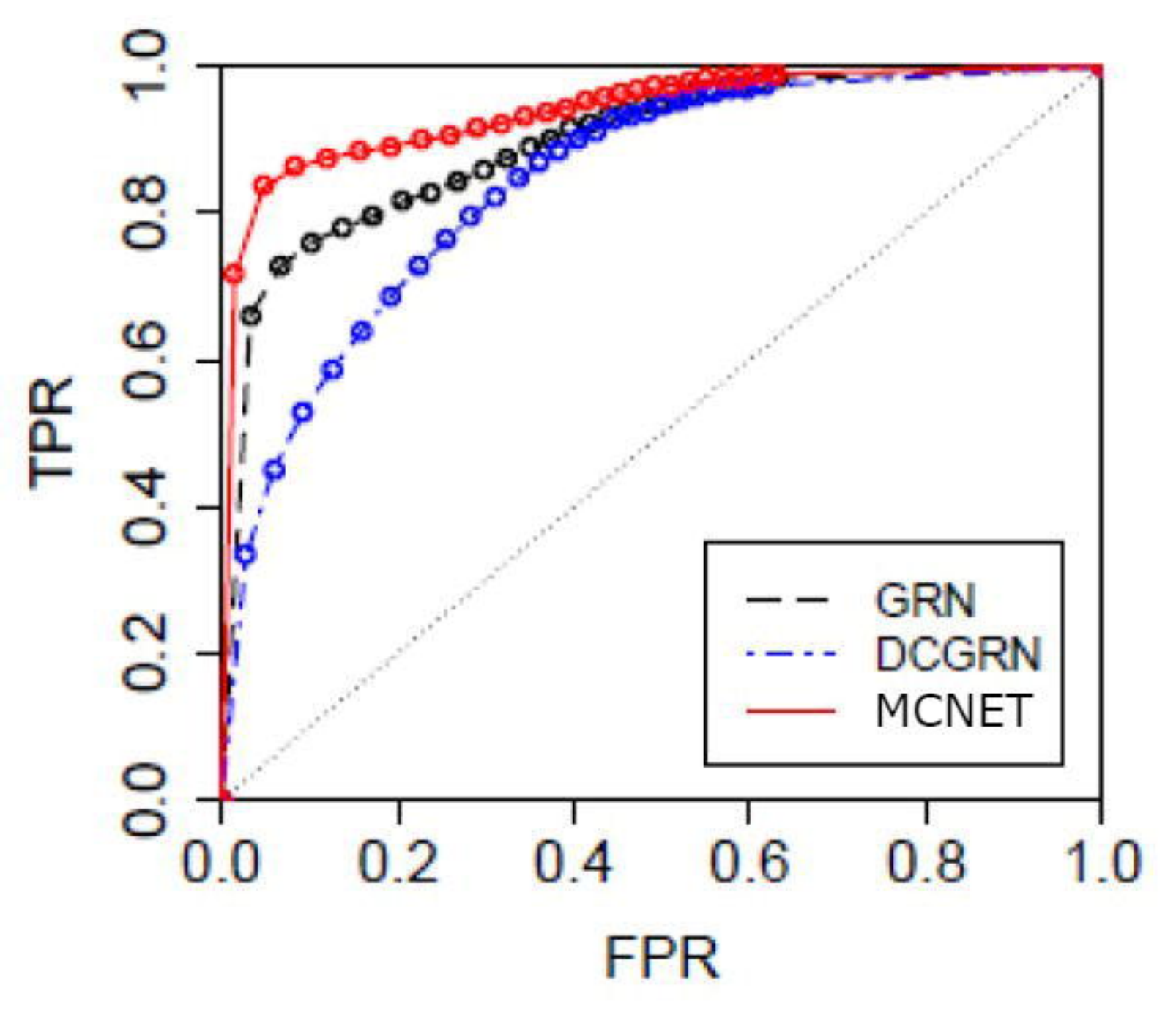

FDR
0.015
0.020
0.025
GRN
DCGRN
MCNET


Sensitivity
0.15
0.25
0.35
MCNET
DCGRN
iGRN


Sensitivity on CNV
0.05
0.10
0.15
DCGRN
MCNET
Sensitivity on DM
0.05
0.10
0.15
DCGRN
MCNET


TPR
0.0
0.2
0.4
0.6
0.8
1.0
20
40
60
80
100
Step on LASSO
DCGRN
MCNET


TPR
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
Sparsity steps
DCGRN
MCNET


TPR
FPR
0.0
0.2
0.4
0.6
0.8
1.0
GRN
DCGRN
MCNET