



## OPEN Understanding complex systems through differential causal networks

Annamaria Defilippo<sup>1</sup>, Federico Manuel Giorgi<sup>2</sup>, Pierangelo Veltri<sup>3</sup> & Pietro Hiram Guzzi<sup>1</sup>✉

In the evolving landscape of data science and computational biology, Causal Networks (CNs) have emerged as a robust framework for modelling causal relationships among elements of complex systems derived from experimental data. CNs can efficiently model causal relationships emerging in a single system while comparing multiple systems, allowing to understand rewiring in different cells, tissues, and physiological states with a deeper perspective. Despite the existence of network models, namely differential networks, that have been used to compare coexpression and correlation structures, causality needs to be introduced in differential analysis to robustly provide direction to the edges of such networks, in order to better understand the flows of information, and also to better intervene in their functioning, for example for agricultural or pharmacological purposes. Resolved to reach this ambitious goal, we introduce Differential Causal Networks (DCNs), a novel framework that represents differences between two existing CNs. A DCN is obtained from experimental data by comparing two CNs, and it is a power tool for highlighting differences in causal relations. After a careful definition and design of DCNs, we test our algorithm to model possible differential causal relationships between genes responsible for the onset of type 2 diabetes mellitus-related pathologies considering patients' sex at the tissue level. DCNs allowed us to shed light on causal differences between sexes across nine tissues. We also compare differences among three possible definitions of DCNs to highlight similarities and differences of biological importance. Code, Data and Supplementary Information are available at <https://github.com/hguzzi/DifferentialCausalNetworks>.

The emergence of network biology and network medicine is changing the way of analysing human diseases<sup>1,2</sup> by introducing a system-level perspective. In such a paradigm, a disease results from the (dysregulated) interaction of multiple and heterogeneous entities and processes that manifest the disease. Consequently, network medicine proposes to model such methods in a single holistic model capable of shedding light on the relations among single elementary building blocks<sup>3,4</sup>.

Network medicine is based on formalism coming from network theory<sup>2</sup>. Thus, biological entities are modelled as graph nodes, while their associations are the edges connecting them. There are various types of networks used to represent different biological scenarios. For example, the PPI network formalism represents proteins as nodes, and the edges represent the biochemical associations among them<sup>5</sup>. In Gene Regulatory Networks (GRN), nodes are genes while the weighted edges model the type of association (e.g. promote, inhibit) calculated as the correlation of their expression<sup>6</sup>. GRNs are usually calculated by looking at the Pearson correlation, Spearman correlation, or mutual information on gene expression in multiple samples<sup>7-9</sup>.

The comparison of biological networks associated with different conditions (e.g., healthy vs. diseases, male vs. female, young vs. older people) is thus a way to find differences in networks representing dysregulated mechanisms. Differential Network Analysis (DNA) algorithms<sup>10,11</sup> have been introduced to analyse such differences. Given two different biological conditions  $C_1, C_2$  represented by two networks  $N_1, N_2$  obtained from experimental observation corresponding to  $C_1, C_2$ , DNA algorithms aim to identify changes in network rewiring that may be associated with changes between phenotypical changes between  $C_1$  and  $C_2$ . Recently, this has been applied to the elucidation of age and sex effects in causing different responses to drugs, outcomes of diseases and incidence of comorbidities for many complex chronic diseases<sup>12-15</sup>.

DNA algorithms are usually applied to protein or gene expression networks built by calculating correlations between the expression profiles of gene/protein pairs<sup>16</sup>. Since correlation does not imply causation, researchers have recently focused on determining a theoretical framework for causality<sup>17,18</sup>. Consequently, some independent

<sup>1</sup>Department of Surgical and Medical Sciences, Magna Graecia University of Catanzaro, Catanzaro 88100, Italy.

<sup>2</sup>Department of Pharmacy and Biotechnology, University of Bologna, 40126 Bologna, Italy. <sup>3</sup>Department of Computer Science, Modelling and Electronics (DIMES), University of Calabria, Rende 87036, Italy. ✉email: [hguzzi@unicz.it](mailto:hguzzi@unicz.it)

works have proposed the introduction of *causal networks*, which explicitly model causality through causal (directed) edges. As described by Pearl<sup>19</sup>, the identification of causal relations between entities can be done using a statistical approach, which can suggest the direction of the edges. Unfortunately, it has been shown that the fields of biology and medicine need more work to determine causal edges since some of the Pearl hypotheses, such as the acyclical nature of the underlying network, do not hold<sup>20</sup>.

While several topological methods exist to assess the differences of gene-gene networks (Schieber et al<sup>21</sup>, Ran et al<sup>22</sup>, Jiang et al<sup>23</sup>), none of them explicitly tackle the *causal* element of such graphs. Our goal is to analyse not only the existence of differences between two causal network but we want to discover the topology and the structure of them (i.e. a qualitative comparison), i.e., the differences among edges, while the presented references find only a quantitative comparison among networks. Our main novelty is that, despite the simplicity of the approach, there are no works formalising and discussing the methodology of Differential Causal Network that we retain should be introduced, similarly to Differential Networks.

*Causality* is a term that represents a generic relationship between a cause and an effect, which is determined by causes. Although the relationship is intuitive, the determination of causality from experimental data is quite a challenge<sup>24</sup>.

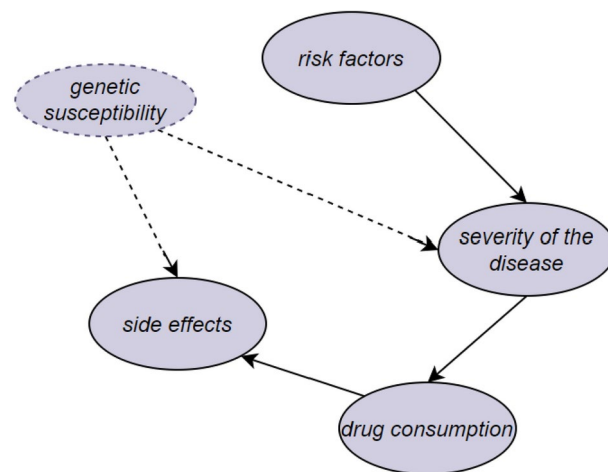
More specifically, determining a causal relationship among a set of variables is equivalent to answering the following questions: (i) How much does a change in a variable alter some other variables (causal inference)? (ii) Which variables do we need to alter in order to modify another variable (causal discovery)? Usually, the variable that causes the change is referred to as treatment, while the variable for which we observe the response is the outcome.

We need a theoretical framework for this analysis to represent and analyse all the elements<sup>25</sup>. The literature contains two main groups of frameworks: namely structural causal models (SCMs)<sup>18</sup> and the potential outcome framework<sup>26</sup>.

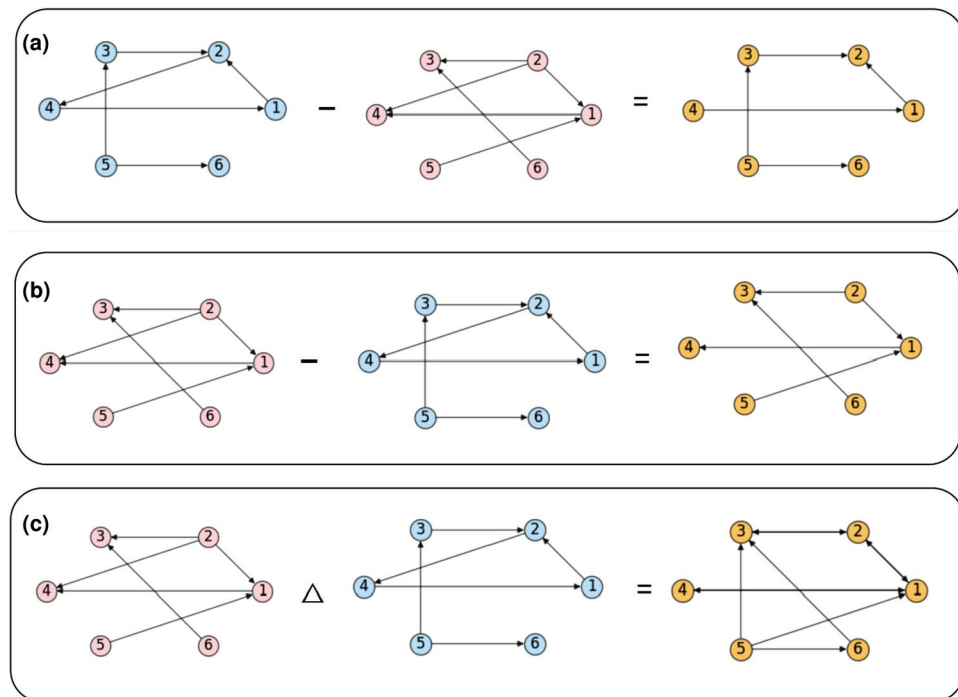
SCMs are based on the use of two components: a causal graph (or causal diagram) that models the relations in a graphical way and the related structural equations between the objects, as represented in Figure 1 where the causal nodes are related to the effects by means of directed edges (also indicated as outgoing edges). Causal graphs are a special case of Bayesian networks modelled as directed graphs. In a causal graph, each node models a random variable (treatment, outcome, observed, and unobserved). A directed edge  $v_i, v_j$  from node  $v_i$  to node  $v_j$  denotes a causal effect of  $v_i$  to  $v_j$ . The structural equation of a causal graph specifies the causal relationships between variables in the system, illustrating how each variable influences and is influenced by others within the graph.

We here aim to bring together both causality and differential approaches, yielding to *Differential Causal Networks*. Given two Causal Networks  $C_1 = \{G_1, E_1\}$  and  $C_2 = \{G_1, E_2\}$  we define a Differential Causal Network  $DCN_{1,2} = \{G_1, E_{1,2}\}$  between the previous two as the network with the same node set and each edge is defined as the difference between the corresponding pair of the edges. In such a way, the adjacency matrix of the DCN is the difference between the adjacency matrices of the two input networks, that is,  $A_{DCN} = A_{C_1} - A_{C_2}$ .

A differential causal network is generated by subtracting one causal network from another or performing symmetric differences between two causal networks, as depicted in Figure 2. It is also worth noting that, being directed graphs, two edges with different directions, even if they connect the same nodes, are considered different edges because their directionality is considered. Taking into account a directed edge  $v_i, v_j$  from node  $v_i$  to node  $v_j$  in  $C_1$  and a directed edge  $v_j, v_i$  from node  $v_j$  to node  $v_i$  in  $C_2$ , performs the following:



**Fig. 1.** In the figure, the element with dashed lines indicates an exogenous variable whose causal impact is unobserved, while the other elements are observed endogenous variables whose connecting arrows define a cause-and-effect relationship. For example, the severity of a disease can be caused by exposure to certain risk factors and, in turn, can be the cause of medication intake. This latter fact is the cause of possible side effects. Furthermore, there may be unobserved factors, such as genetic susceptibility, that influence both the severity of the disease and the adverse effects of the drug.



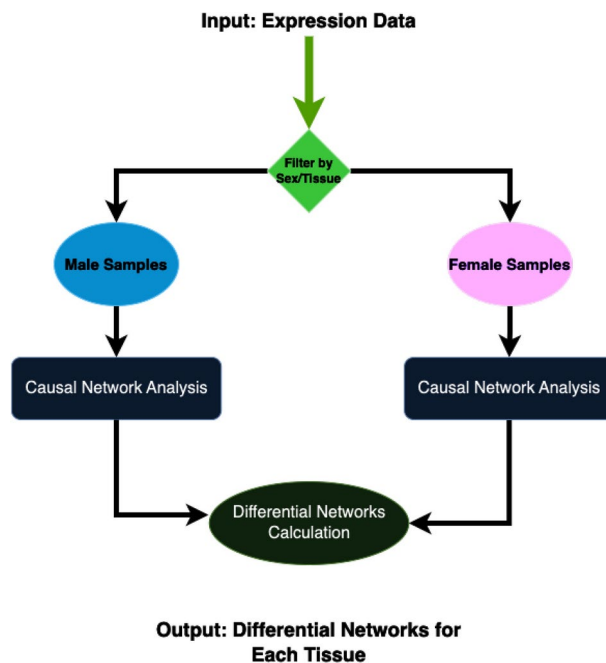
**Fig. 2.** The figure shows the three possible ways to obtain a difference network starting from two input graphs. **(a)** Starting from two input graphs (represented by using blue and pink nodes), the differential network (represented by using yellow edges) contains only the edges present in the blue graph but not present in the pink graph. **(b)** Starting from two input graphs (represented by using blue and pink nodes), the differential network (represented by using yellow edges) contains only the edges present in the pink graph but not present in the blue graph. **(c)** Starting from two input graphs (represented by using blue and pink nodes), the differential network (represented by using yellow edges) contains only the edges present in only one of the input graphs.

- the symmetric difference between  $C_1$  and  $C_2$ , the adjacency matrix will contain a value equal to 1 in both the position  $i, j$  and the position  $j, i$ ;
  - the difference between  $C_1$  and  $C_2$ , the adjacency matrix will contain a value equal to 1 in the position  $i, j$  if the edge  $v_i, v_j$  is not in  $C_2$ ;
  - the difference between  $C_2$  and  $C_1$ , the adjacency matrix will contain a value equal to 1 in the position  $j, i$  if the edge  $v_j, v_i$  is not in  $C_1$ ;
- Differential causal networks represent the changes in causality relations among nodes between two conditions or states. Nodes changing their connections between two states are then considered to be rewired and are biologically meaningful. As in Differential Networks where nodes involved in change of state rewire more frequently w.r.t. those that are not, in our proposal the rewired nodes are used to identify subnetworks related to change of states. Similarly to Differential Networks, since nodes that are related to the change of state rewire more frequently than nodes that are not involved in the change of state, the rewired nodes may be used to identify subnetworks responsible for the insurgence of different states.

We report our definition of DCN and then we report how it can be used to represent and analyse a biological scenario following the pipeline described in Figure 3. We analyse data related to Type 2 Diabetes Mellitus (T2DM) which is a complex disease presenting differences in comorbidities and incidence due to age and sex<sup>12,27,28</sup>.

For instance, Guzzi et al<sup>28</sup>. investigated differential networks at the tissue level to understand phenotypical differences in diabetes comorbidities between males and females, without considering causal relationships between genes. Differential Network Analysis (DNA) algorithms derive differences in network structures between two conditions by analyzing experimental gene-expression data, using statistical methods to identify changes in edges and nodes of co-expression networks. Unlike existing works, a Differential Causal Network (DCN) investigates the difference of causal edges between two causal networks composed of the same nodes, taking into account link directions.

Here, we hypothesise the existence of differences even in causal mechanisms. We start from gene expression data stored in the GTEx database<sup>29</sup> and comorbidities of TD2M stored in the T2DiACOD database. Focusing on T2DM, we obtained results from tissues organised into six age groups: 20-29, 30-39, 40-49, 50-59, 60-69, and 70-79 years. We obtained different samples for each tissue. We subdivided samples considering sex, and we derived CNs for males and females and a representative DCN for each tissue. Our results show that some DCNs may explain differences between sex.



**Fig. 3.** Figure reports the pipeline of the proposed use of DCN to compare expression gene data using causal network analysis. We start by considering the gene expression of males and females. First, we separate the groups, and we apply a causal discovery algorithm for building a causal network. Corresponding pairs of causal networks are then compared to obtain a single Differential Causal Network. The process is repeated for each considered tissue.

The proposed use of DCNs can be considered a novelty in the use of causal networks inferred from the largest human transcriptomics dataset. Indeed, the provided implemented solution, available on a GitHub repository, to the best of our knowledge is the first example of definition and application of differential causality analysis in a dataset with respect to sex, where biological pathways enriched in differential male vs. female causal gene networks are also analysed and reported. Furthermore, our work shows the following main novelties:

1. Causality is inferred from the largest human transcriptomics dataset, based on the number of samples.
2. First implementation of differential causality analysis in the largest dataset comparing male and female samples.
3. Critical evaluation of biological pathways enriched in differential male vs. female causal gene networks.

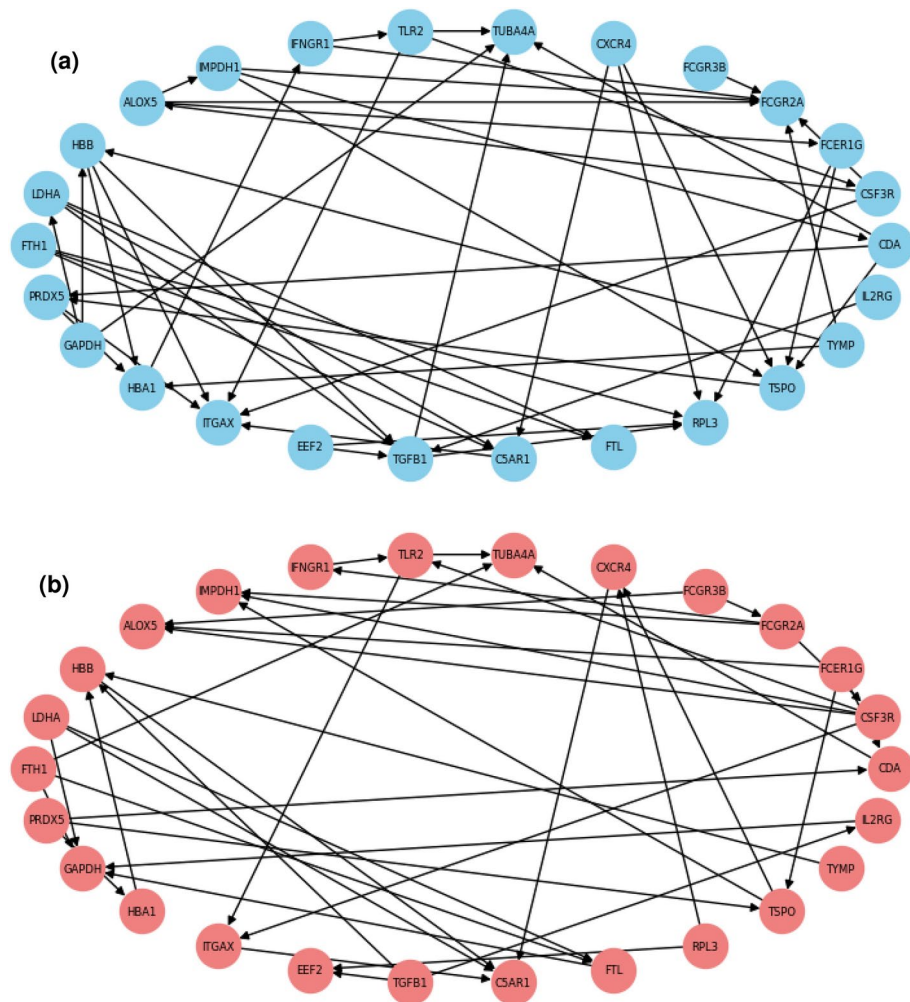
### A toy example

To demonstrate the feasibility of applying differential causal networks in a biological context, such as gene expression, we present a toy example. We start by deriving two causal networks obtained by using the PC algorithm in two gene expression datasets<sup>30,31</sup>. We split the dataset by considering the sex of the sample donor and we generate two networks, one for men and one for women, as depicted in the Figures 4a and 4b.

Once two networks were obtained the differential causal network, illustrated in the following figures, were calculated. Three ways of showing the differences, as described previously, were performed and plotted. The Figure 5a illustrates the differential causal network obtained by applying the symmetric difference between the two causal networks of men and women. It contains the edges not common to the two networks, meaning those present are in either the male or only in the female causal network. In the Figure 5b, the differential causal network between the male and female causal networks is depicted, containing only the edges present in the former that are not present in the latter. Similarly, in the figure 5c, the differential causal network is computed as the difference between the female and male networks, containing only the edges present in the female network that are not present in the male one.

### Synthetic dataset test

To prove that the methodology is “robust” and only produces results when the initial networks compared are truly different, we performed some experiments on synthetic data using three pairs of datasets with the same pairwise causal relations and we found three empty DCNs for each pair. We tried our methodology on three synthetic datasets with increasing number of fictitious genes (26 genes, 52 genes and 104 genes), obtaining the results reported in Table 1. We generated the first synthetic dataset using an ad hoc generated script. Next, we expanded the first pair of datasets by generating a new column for each pre-existing column to which a random number was added. In this way, we generated the second synthetic pair of datasets with a double number of genes compared to the second and the third synthetic pair of datasets with double the number of genes compared to the second. The PC algorithm derived identical causal networks, as expected, since the same causal



**Fig. 4.** Causal Network obtained for males and females, including 26 genes pharmacologically active, which are expressed in the blood tissue. **(a)** Male causal network **(b)** Female causal network.

mechanisms persisted in the data. Figure 6 depicts the causal networks obtained with 26 (6a), 52(6b) and 104 genes(6c). Consequently, the DCNs computed with the three proposed differences were empty.

### Related work

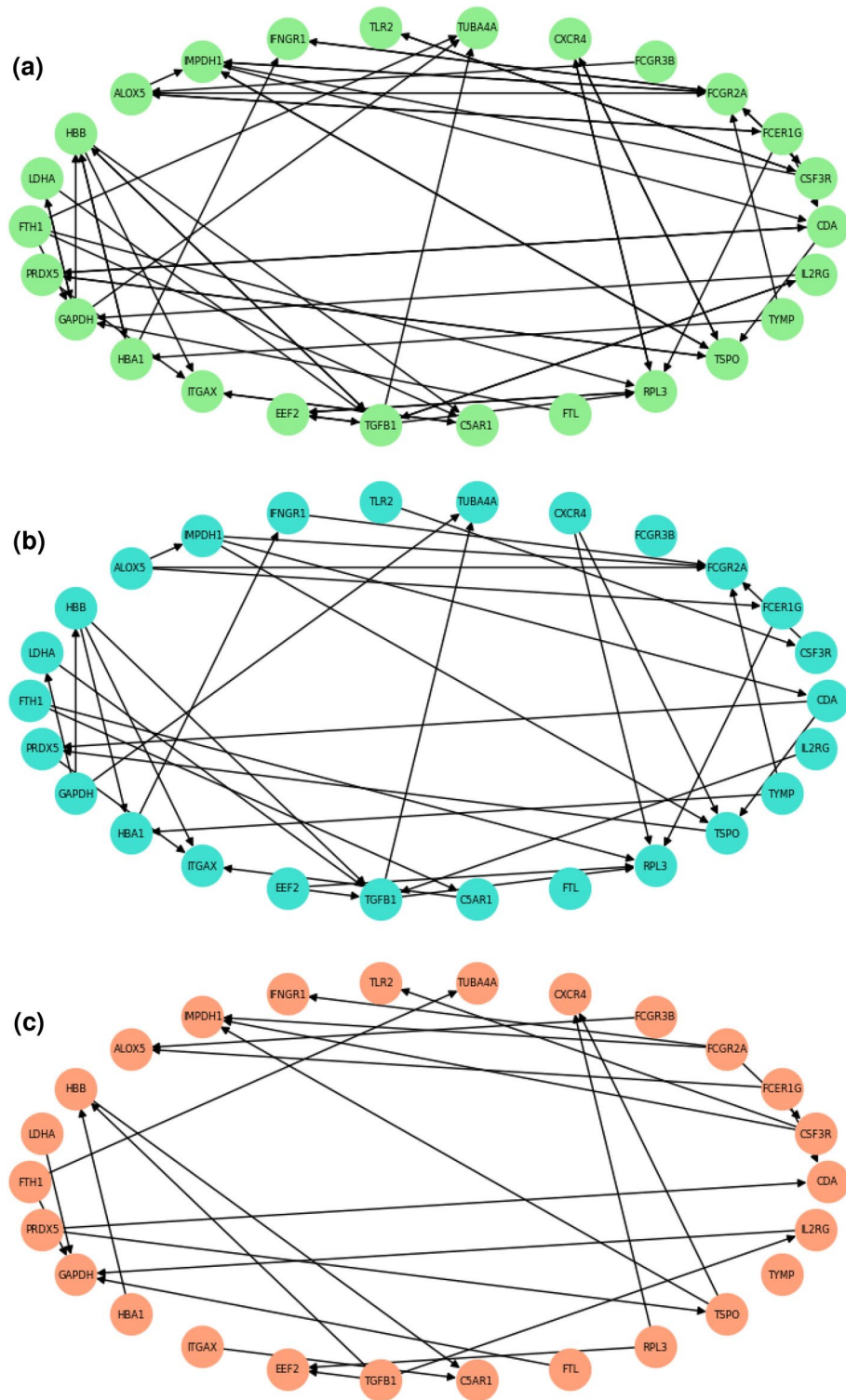
Causal structure learning or causal discovery includes some methods aimed at finding the causal graph from the structure of the data-generating process. Considering traditional causal discovery, the causal structure is recovered from observational data only<sup>31</sup>. However, more recent approaches allow the incorporation of expert knowledge into the graphical representation from interventional data. Therefore, causal discovery algorithms aim to estimate the causal structure of the mechanism that generated the data passed to the algorithm. According to<sup>18,31</sup>, causal discovery methods may be divided into 4 classes: (i) Constraint-based methods, (ii) Score-based, (iii) Functional causal discovery, and (iv) Gradient-based methods.

In our analysis, we used the PC algorithm (Peter-Clark algorithm)<sup>32</sup>, a successor of the IC algorithm, proposed by Verma and Pearl in 1990<sup>33,34</sup>. The PC algorithm aims at identifying the causal structure of a set of variables from observational data. The algorithm belongs to the first class of causal structure learning, based on conditional independence tests. This class focuses on performing a sequence of hypothesis tests to discover the causal graph based on the independence conditions inferred from the data<sup>35</sup>. PC algorithm works by performing a series of conditional independence tests to determine the relationships between variables, ultimately constructing a directed network that represents these causal relationships. This algorithm may be resumed in the next steps<sup>30,31</sup>:

1. a complete undirected graph is created and a succession of conditional independent tests are performed for all pairs of variables:

- elimination of edges between two unconditionally independent variables;
- elimination of edges between two conditionally independent variables given a third variable connected to either of them;





**Fig. 5.** Differential Causal Networks obtained by taking the difference between male and female networks. **(a)** Symmetric difference between the male and female networks. **(b)** Difference between the male and female networks. **(c)** Difference between the female and male networks.

- elimination of edges between two conditionally independent variables given a couple of other two variables connected by an edges between them and connected to one or both the first two considered variables;
- repetition of the last step checking conditional independencies on a subset of  $n$  variables (increasing  $n$ ), while there are no more couple of variables adjacent to one or both of the initial variables;

2. checking  $v$ -structures to orient the edges;

Case	Number of causal edges				
	Causal network A	Causal network B	DCN A-B	DCN B-A	Symmetric DCN
Case 1: 26 genes	27	27	0	0	0
Case 2: 52 genes	60	60	0	0	0
Case 3: 104 genes	145	145	0	0	0

**Table 1.** Experiments on synthetic data.

3. applying *orientation propagation* to logically direct the edges on the basis of learned information. When the conditional independence decisions are correct, the PC algorithm converges to the Markov Equivalence Class (MAC), considering the Causal Markov and Faithfulness assumptions, i.i.d. samples, and no unmeasured confounders<sup>30</sup>. Considered that, there are situations where a causal discovery method, such as PC, returns a complete partially directed acyclic graph (CPDAG). Therefore it is appropriate to clarify that a CPDAG uniquely represents the corresponding Markov equivalence class. In a CDPAAG, a directed edge stands for an edge that is present in all DAGs of the MAC, while an undirected edge plays a role of an edge whose direction cannot be attributed as it is not unique in the MAC<sup>31,36</sup>. Given the aim of causal discovery methods, these algorithms have become popular for the analysis of genetic mechanisms<sup>36</sup> in the attempt to find new knowledge regarding their regulation and mutual influence, especially for specific diseases.

## Results

Identifying the differential causal networks between different sex may give insights into the molecular mechanisms explaining different incidences and prognoses of comorbidities of pathologies. We here try to shed light on causal mechanism related to comorbidities of Type 2 Mellitus Diabetes as testing framework for the proposed formalism. Consequently we aim to produce and evaluate differential causal networks between males and females using gene expression datasets from GTEx database for adipose, amigdala, aorta, blood, bolon, cerebellum, coronary, liver, and lung. We focused our analysis on the genes related to comorbidities of diabetes as reported in T2DiACoD database<sup>28</sup>. We obtain three DCNs for each tissue considering respectively: Male minus Female (M-F), Female minus Male (F-M) and Symmetric Difference (SD). For each network, we analysed the pathway enrichment considering the KEGG pathway database<sup>37–40</sup> by using the STRING enrichment app<sup>41</sup> of the Cytoscape software<sup>42</sup>(version 3.10.1, <https://cytoscape.org/>). Pathway Enrichment Analysis allows to extract pathway-level information from a large list of genes, coalescing individual contributions in differential (causal) expression and determining which areas of cellular activity are most affected by the investigated variable (in our study, sex). Using STRING enrichment app of the Cytoscape software, the False Discovery Rate (FDR), defined as the proportion of false positive classifications (false discoveries) among all positive classifications (rejections of the null hypothesis), is computed using the Bonferroni method to assess the statistical significance of the enrichment results. Indeed, we selected the top pathways with the lowest FDR for each considered DCN.

Moreover we also compare differences among the three ways of obtaining Networks by considering their similarities measures as the Jaccard Index of the edge set and of the set of pathways.

When deriving causal networks, there is the probability of obtaining some undirected edges that imply that some causal relations among variables can not be read from the graph directly, so there is no confirmed evidence of causal. In our paper, the number of undirected edges are a tiny fraction of the total edges obtained. For instance, for adipose tissue we obtained 2 undirected edges out of 989 edges in male network and 1 undirected edge out of 799 edges in female network. We report all the data about undirected edges in Table 2. It's important to clarify that 'undirected' edges are those without a defined direction, representing a bidirectional relationship. We did not remove these edges because a bidirectional edge, for example, in the male network but not in the female network, is a difference to take into account. This relationship identified in one network may not be identified in the other as it does not constitute such relevant causality, opening the possibility of further investigation.

Furthermore, we should note that despite the initial datasets presents the same number of genes in all the cases, the generated network present a remarkable difference both among tissues and, more important, even between the network of the same tissue corresponding to different sex as reported in Figures 7, 8, and 9. The proportion of edges in various tissues remains the same between men and women as it is possible to note in the Figure 7. Despite this similarity, the number of edges is quite different among sexes; in fact, in male networks, comparing with female networks, a greater number of edges in all tissues is exhibited.

Additionally, the Figure 8 illustrates the number of edges in the differential causal networks, while the Figure 9 shows symmetric differential causal network. In summary, it can be argued that there is not only a numerical difference between the networks, but also a difference in causal connections between nodes.

In order to provide also a qualitative comparison among the networks, we evaluated the Jaccard index using edges, i.e. the way nodes are connected. We use the Jaccard Index to compare the graphs. The Jaccard index is a measure of the similarity between two sets, that is defined as the size of their intersection divided by the size of their union. In graph theory, the Jaccard index is hence used to measure the similarity between the sets of the neighbors of two nodes. Given two nodes  $u$  and  $v$  in a graph  $G$ , the Jaccard similarity  $J(u,v)$  is defined as the ratio between the shared neighbours and the union of the set of neighbors. In our definition we extend this definition and we consider the number of shared edges. Since we manage ordered graphs, we also consider the direction of the edges when comparing them and we average the Jaccard Index of all the node pairs. Starting from the the general formula:

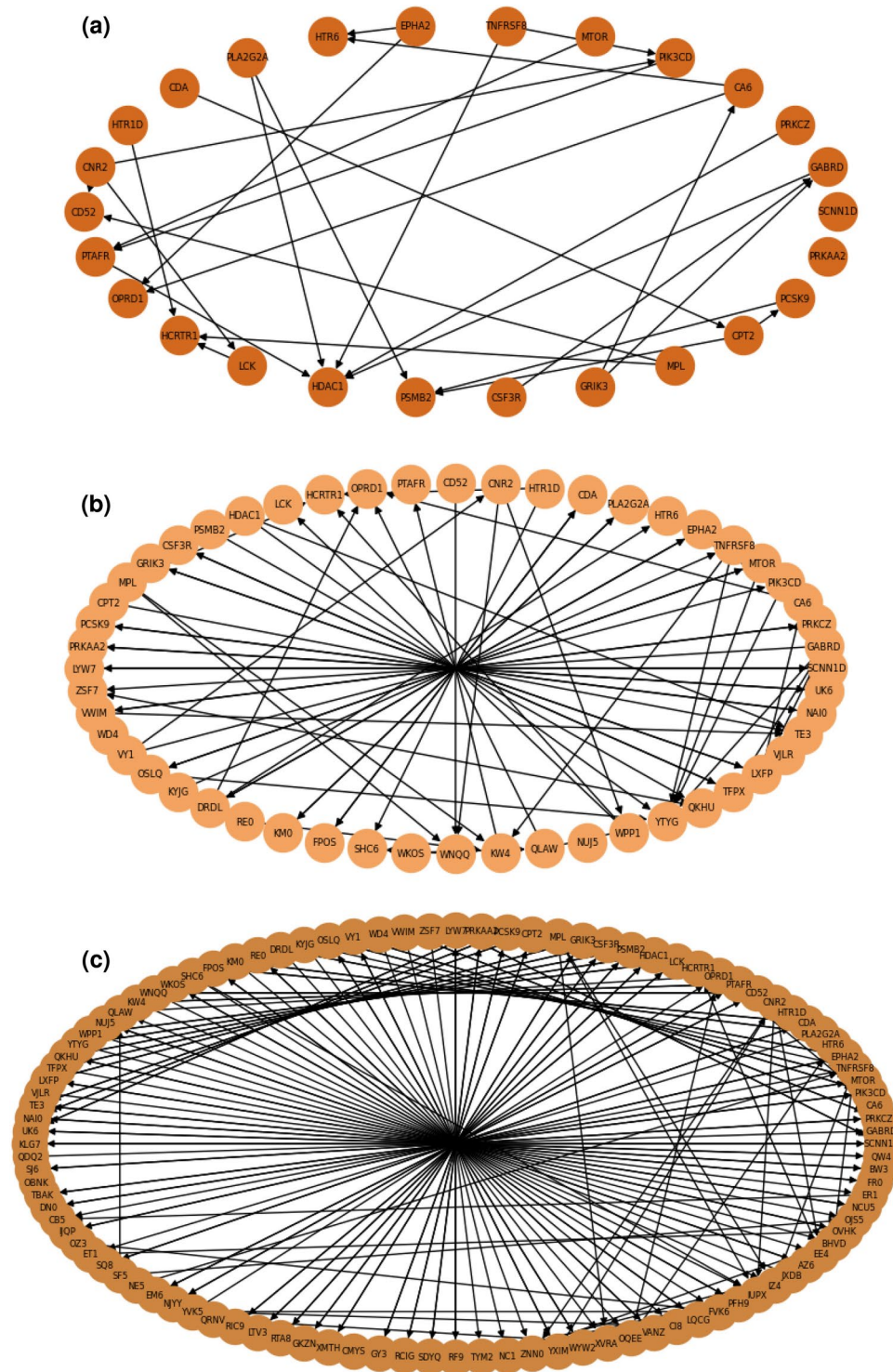
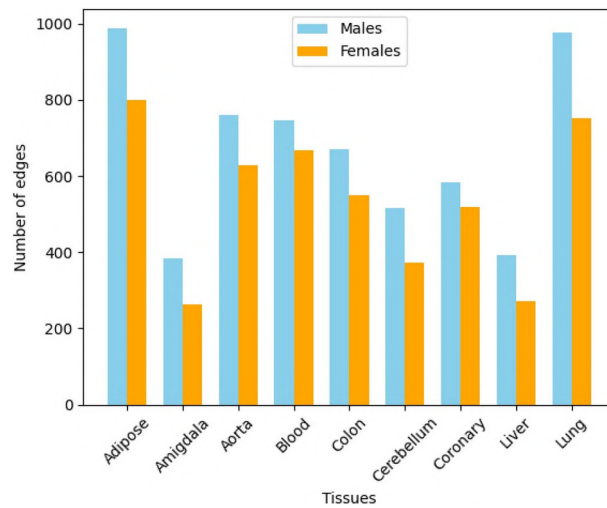


Fig. 6. Causal Networks obtained by using PC on synthetic datasets.

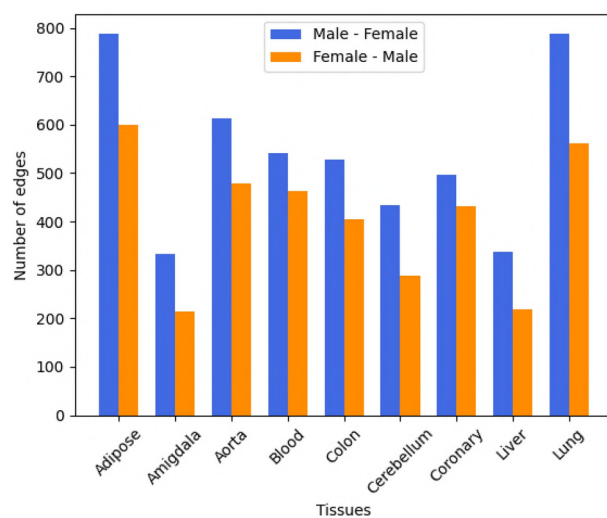


Tissue	Number of undirected edges	
	Male network	Female network
Adipose tissue	2	1
amigdala tissue	9	8
Aorta tissue	2	5
Blood tissue	6	8
Colon tissue	1	1
Cerebellum tissue	3	15
Coronary tissue	2	3
Liver tissue	4	6
Lung tissue	2	2

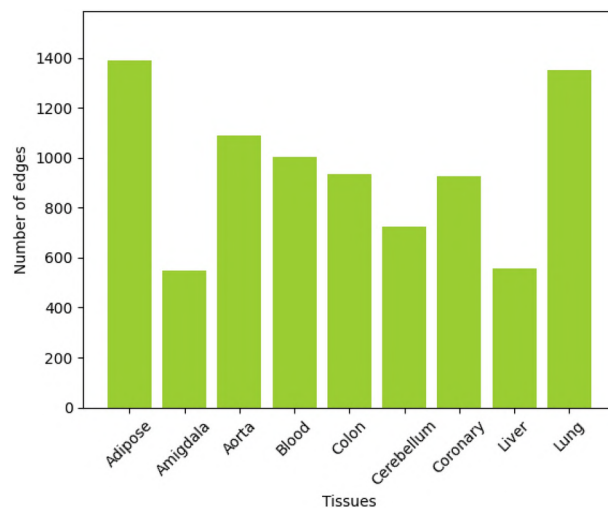
**Table 2.** Number of undirected edges in CPDAGs in male and female causal networks for each tissue.



**Fig. 7.** A graphical comparison is provided to illustrate the number of edges obtained for the nine tissues under examination in the context of the male and female causal networks.



**Fig. 8.** The number of edges obtained for the nine tissues under examination in the context of the male and female differential causal networks are depicted to point out the differences. In blue, the number of edges in males minus females differential causal networks is presented; in orange, the number of edges in females minus males differential causal networks.



**Fig. 9.** To complete the edge comparisons, the number of edges obtained for the analyzed tissues in the symmetric differential causal networks. The number is visibly higher than the previous cases being the result of exclusive disjunction.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

where A, B are two sets of items, e.g. nodes or edges for graphs. We propose a modified Jaccard index as follows:

$$J(G_1, G_2) = \frac{1}{|N|} \sum_{n \in N} \frac{|E_{G_1}(n) \cap E_{G_2}(n)|}{|E_{G_1}(n) \cup E_{G_2}(n)|}$$

where:

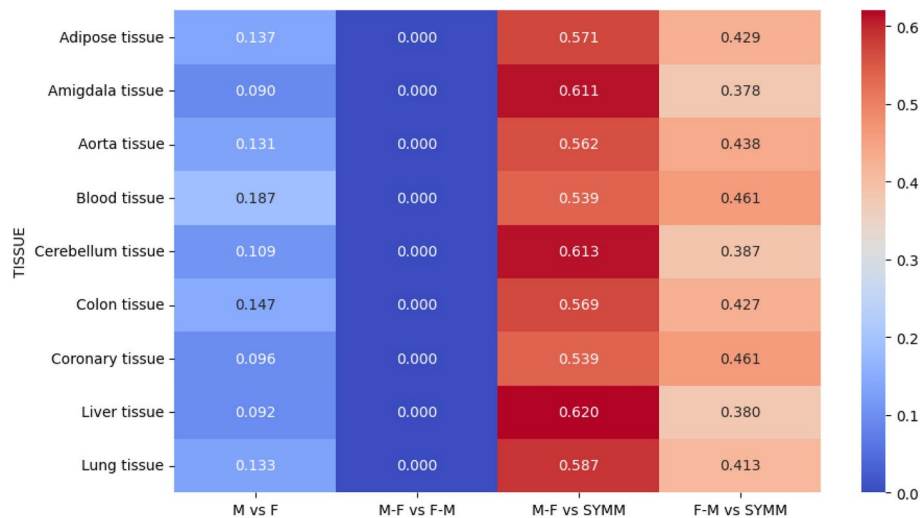
- N is the set of nodes in the graphs (considering  $G_1$  and  $G_2$  having the same nodes)
- $E(n)$   $E_G(n)$  indicates the set of edges incident to the node  $n$  are the edges of the node  $n$  of the graph; Jaccard indexes for each pairwise comparison between differential causal networks are reported in the Figure 10.

According to Jaccard similarity's meaning, 0 suggests a lack of similarity, while 1 denotes maximum similarity. Given this, second column of the Figure 10 indicates the complete absence of similarity between the differential causal networks computed performing the difference *male causal network - female causal network* and the difference *female causal network - male causal network*.

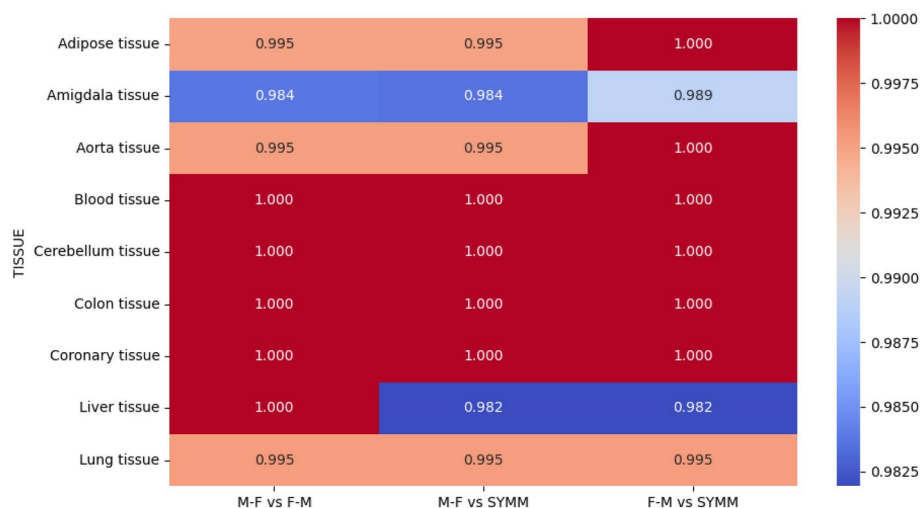
Jaccard Index was computed also between the differential causal network of *male causal network - female causal network* and the symmetric differential network and between the differential causal network of *female causal network - male causal network* and the symmetric differential network. As expected, according to the Jaccard index, there is similarity in the latter cases. However, the differential causal network *male causal network - female causal network* results more similar to the symmetric differential network than the differential causal network *female causal network - male causal network*. The dimensions of the DCN are determined by two factors: the dimensions of the networks of the conditions we compare and the number of differences between them. The dimensions of the compared networks are due to the intrinsic number of causal relationships and the characteristics of the causal discovery algorithm. Regarding the second factor, it should be noted that while a larger number of links may yield a larger number of differences, the number of different links definitely depends on the number of different links, which may not be directly proportional. In our specific case, the more significant similarity may be explained by the fact that male causal networks showed more links. Nevertheless, this could be a direct consequence of the causal discovery algorithm, which identified a greater number of causal relationships between genes, taking into account gene expression in the male sex.

As a final comparison at this phase, Jaccard Index was calculated between male causal networks and female causal networks whose values are reported in Figure 10. The results obtained, being all close to 0, further confirm the difference between the two causal networks already highlighted by the differential causal networks and their analysis.

Performing pathway enrichment, we compared, using Jaccard index, the KEGG Pathways obtained in pairs between the DCNs, as illustrated in the Figure 11. We performed the computation of Jaccard Index applying general formula, in which sets were composed by the unique description of the pathways. Contrary to expectations, there is a high similarity between the various enrichment pathways in the cases considered.



**Fig. 10.** Heatmap for Jaccard index to graphically visualise and underline the obtained results. The first column is the similarity among Causal Networks of Males vs Causal Networks of Females. The second column is the Jaccard Index between the edge set of the DCN calculated as male minus female compared with the DCN of female minus male. It reveals that the calculation of a DCN is not symmetric. While the third and fourth columns are the Jaccard Index between Males minus Females with respect to the Symmetric DCN and Females minus Males with Symmetric DCN. Networks calculated as Males minus Females have a greater similarity due to the higher number of edges.



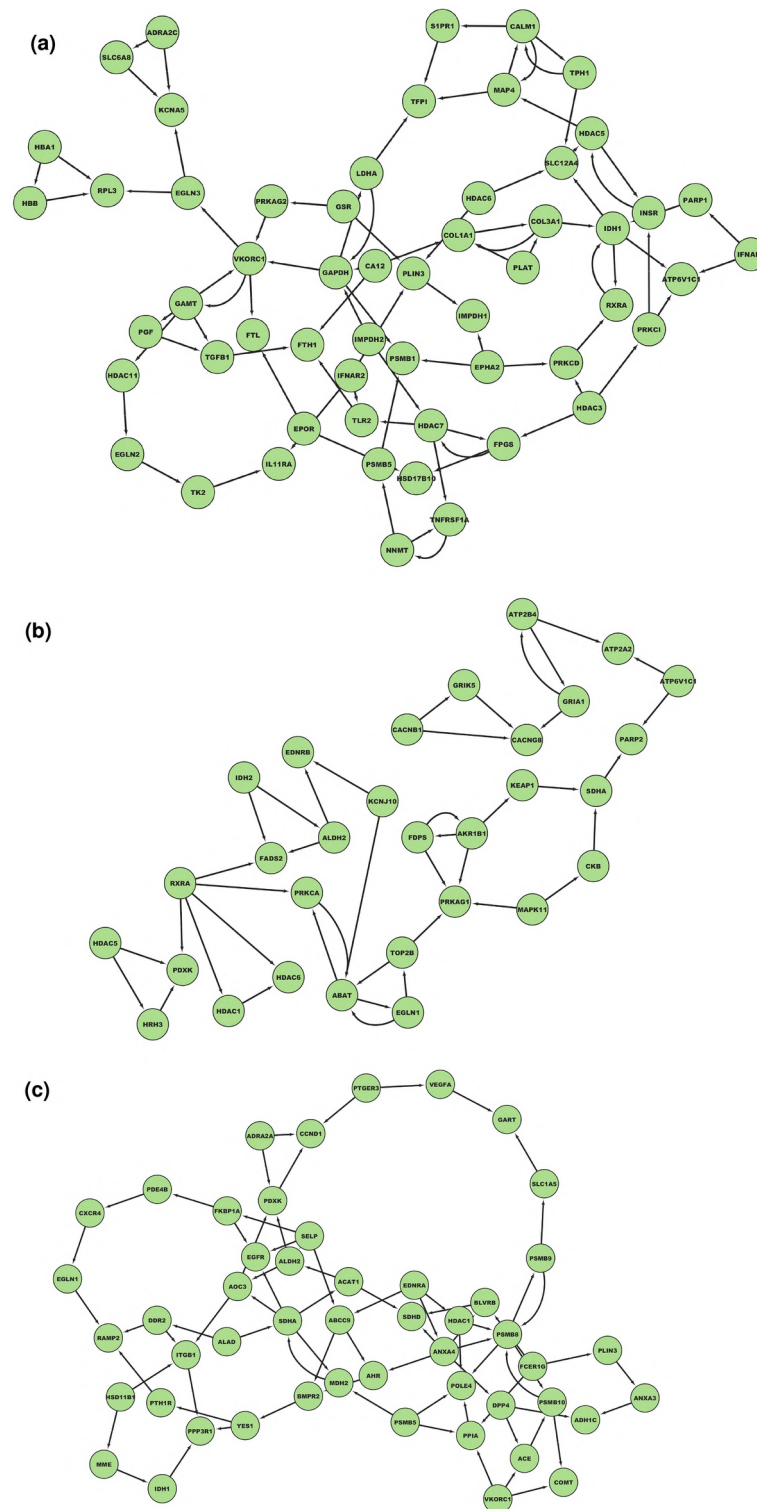
**Fig. 11.** Heatmap for Jaccard index to graphically visualize the comparison between pairs of DCNs compared with respect to enriched pathways. The first column is the similarity between DCN M-F enriched pathway compared to DCN F-M enriched pathway. The second column is the Jaccard Index between DCN M-F enriched pathway compared to Symmetric DCN enriched pathway. While the third column depicts the Jaccard Index between DCN F-M enriched pathway compared to Symmetric DCN enriched pathway.

However, slightly greater differences than others, although small, were observed in the enrichment pathways of amygdala and liver tissue compared to other cases.

In the following subsections we reported the DCNs obtained using the simple difference between causal networks (M-F and F-M) and using the symmetric difference between causal networks. For a more effective visualization of networks that are rather large in some cases, clustering was performed on each network, and the largest identified cluster was included to enhance network visibility. Consequently, “clustered” in this context refers to the graphical layout and visualization of the selected networks. For the latter aim, the largest clusters calculated using MCODE of the Cytoscape software with default parameters were reported. The only two exceptions, amygdala tissue and cerebellum tissue for the F-M DCNs, for which, respectively, a Degree cutoff of 5 and 4 was set.

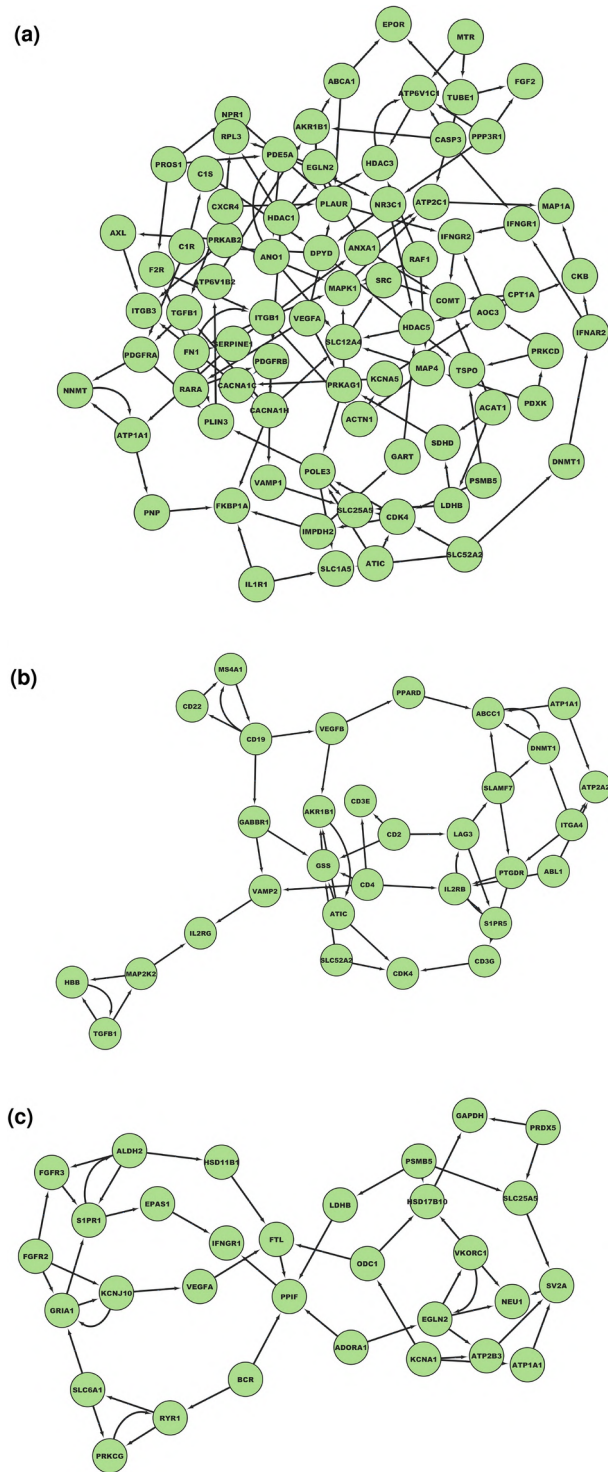
## Differential networks obtained with symmetric difference

Symmetric Differential Networks have been obtained performing symmetric difference for the selected tissues, and their representations are displayed in Figures 12, 13, 14 depicts these networks. For each symmetric differential network, we also detected the enriched pathways of which the tops with the lowest FDR values are reported in Tables 3,4,5,6,7,8,9,10,11.



**Fig. 12.** Symmetric DCNs (1): for a better visualization, we represent only a fragment of the networks. The whole networks are available as supplementary material. **(a)** SYMMETRIC DCN in the aorta tissue. **(b)** SYMMETRIC DCN in amigala tissue. **(c)** SYMMETRIC DCN in adipose tissue.

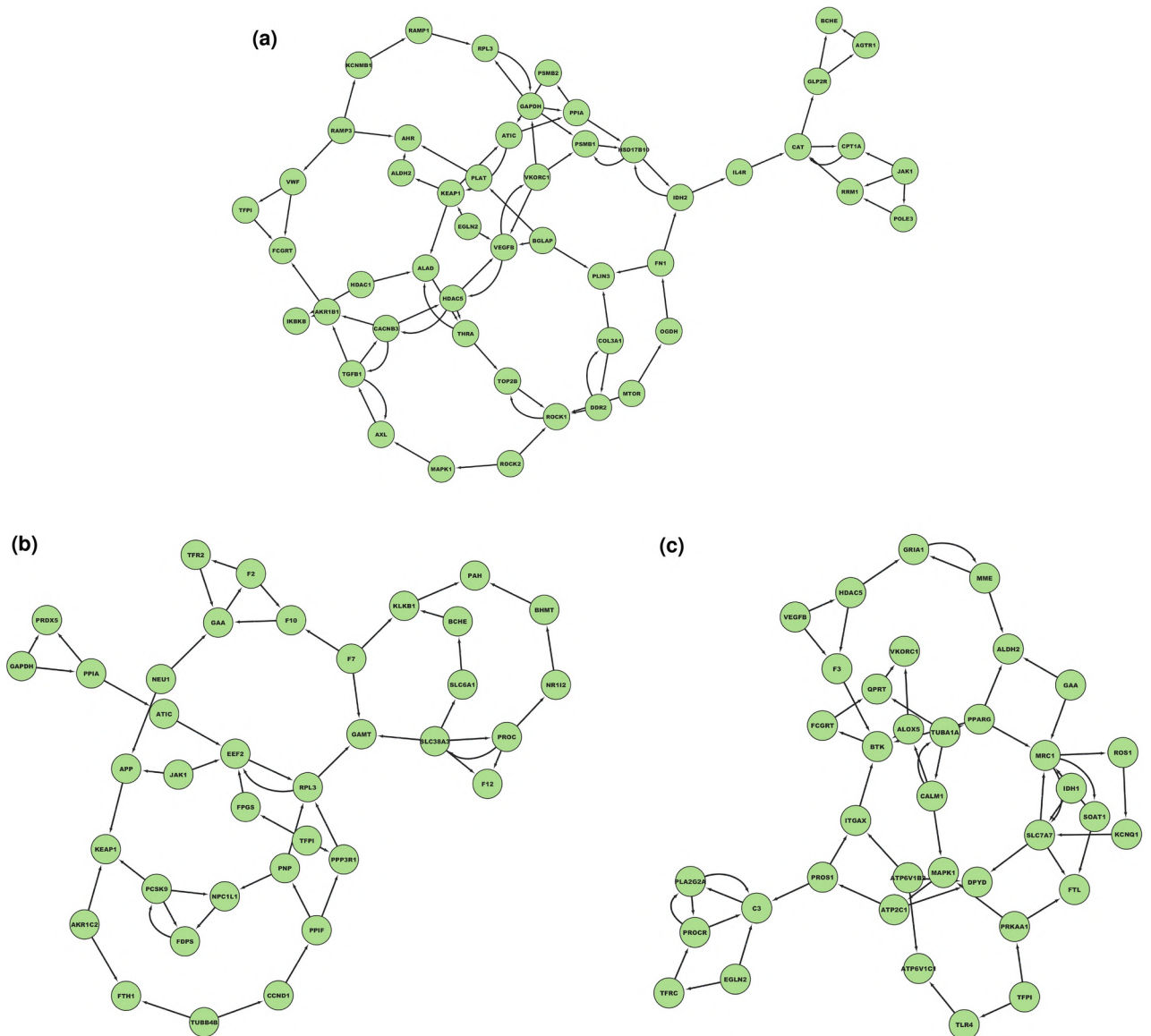




**Fig. 13.** Symmetric DCNs (2): for a better visualization, we represent only a fragment of the networks. The whole networks are available as supplementary material. **(a)** SYMMETRIC DCN in coronary tissue. **(b)** SYMMETRIC DCN in blood tissue. **(c)** SYMMETRIC DCN in cerebellum tissue.

### Differential networks obtained as M-F

This section shows the differential networks obtained as Male minus Females in the selected tissues. Figures 15, 16 depict these networks. For each network we also found the enriched pathways as shown in Tables 12,13,14,15,16,17,18,19,20. The summarized results from the Male-Female (M-F) Differential Network analysis in various tissues reveals significant enrichment of key biological pathways associated with signal transduction, disease mechanisms, and metabolic processes. These findings underscore the intricate molecular interplay underlying differential disease susceptibilities and responses between males and females. Highlighted pathways



**Fig. 14.** Symmetric DCNs (3): for a better visualization, we represent only a fragment of the networks. The whole networks are available as supplementary material. (a) SYMMETRIC DCN in colon tissue. (b) SYMMETRIC DCN in liver tissue. (c) SYMMETRIC DCN in lung tissue.

Pathway	FDR value
PI3K-Akt signalling pathway	2.09E-29
MAPK signalling pathway	2.48E-18
cGMP-PKG signalling pathway	1.69E-17
HIF-1 signalling pathway	8.24E-17
Rap1 signalling pathway	2.35E-16
AGE-RAGE signalling pathway in diabetic complications	2.38E-16
Proteoglycans in cancer	6.04E-16
Osteoclast differentiation	1.23E-15

**Table 3.** Top Enriched KEGG Pathways in Adipose in Symmetric Differential Network.

Pathway	FDR value
Alzheimer disease	2.79E-13
MAPK signalling pathway	7.72E-13
Oxytocin signalling pathway	2.34E-12
Thyroid hormone signalling pathway	1.5E-11
cAMP signalling pathway	4.53E-10
Adrenergic signalling in cardiomyocytes	1.72E-09
Rap1 signalling pathway	1.87E-09
Gap junction	3.36E-09

**Table 4.** Top Enriched KEGG Pathways in amigdala in Symmetric Differential Network.

Pathway	FDR value
PI3K-Akt signalling pathway	5.75E-20
Proteoglycans in cancer	5.64E-14
MAPK signalling pathway	9.65E-14
cGMP-PKG signalling pathway	9.65E-14
Osteoclast differentiation	2.42E-13
FoxO signalling pathway	5.25E-13
AGE-RAGE signalling pathway in diabetic complications	5.86E-13
Influenza A	5.86E-13

**Table 5.** Top Enriched KEGG Pathways in Aorta in Symmetric Differential Network.

Pathway	FDR value
PD-L1 expression and PD-1 checkpoint pathway in cancer	8.87E-22
Human T-cell leukemia virus 1 infection	2.74E-20
Leishmaniasis	1.82E-16
Th1 and Th2 cell differentiation	3.94E-16
Tuberculosis	7.31E-16
Natural killer cell mediated cytotoxicity	1.22E-14
Influenza A	5.1E-16
T cell receptor signalling pathway	5.5E-16

**Table 6.** Top Enriched KEGG Pathways in Blood in Symmetric Differential Network.

Pathway	FDR value
Oxytocin signalling pathway	2.04E-17
MAPK signalling pathway	4.44E-16
Alzheimer disease	1.15E-15
cGMP-PKG signalling pathway	1.34E-11
cAMP signalling pathway	1.34E-11
Prion disease	8.38E-10
Adrenergic signalling in cardiomyocytes	1.66E-09
GABAergic synapse	1.66E-09
Serotonergic synapse	2.38E-09

**Table 7.** Top Enriched KEGG Pathways in Cerebellum in Symmetric Differential Network.

Pathway	FDR value
PI3K-Akt signalling pathway	5.32E-17
cGMP-PKG signalling pathway	1.01E-15
Oxytocin signalling pathway	1.01E-15
MAPK signalling pathway	4.02E-15
MicroRNAs in cancer	4.31E-14
Alzheimer disease	1.01E-13
EGFR tyrosine kinase inhibitor resistance	5.15E-13
Central carbon metabolism in cancer	1.26E-12

**Table 8.** Top Enriched KEGG Pathways in Colon in Symmetric Differential Network.

Pathway	FDR value
PI3K-Akt signalling pathway	1.81E-22
MAPK signalling pathway	9.83E-17
Osteoclast differentiation	1.63E-14
Proteoglycans in cancer	1.95E-14
FoxO signalling pathway	2.42E-14
AGE-RAGE signalling pathway in diabetic complications	2.47E-14
cGMP-PKG signalling pathway	2.59E-14
Th17 cell differentiation	3.42E-14

**Table 9.** Top Enriched KEGG Pathways in Coronary in Symmetric Differential Network.

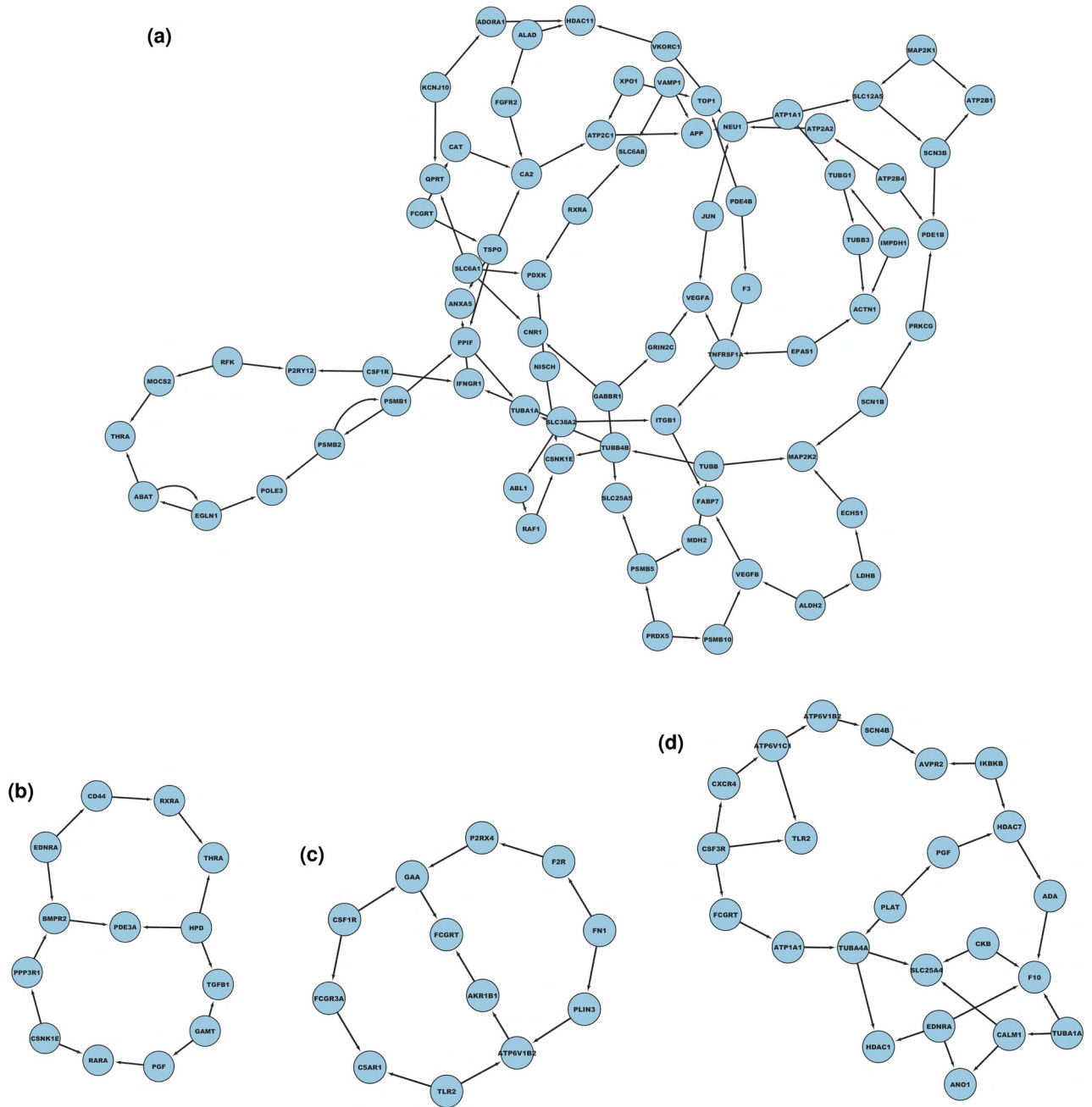
Pathway	FDR value
Complement and coagulation cascades	4.67E-25
Carbon metabolism	0.00000322
Alzheimer disease	0.0000191
Tyrosine metabolism	0.0000336
Th17 cell differentiation	0.0000365
Fatty acid degradation	0.0000882
PI3K-Akt signalling pathway	0.000288
Non-alcoholic fatty liver disease	0.00102

**Table 10.** Top Enriched KEGG Pathways in Liver in Symmetric Differential Network.

Pathway	FDR value
PI3K-Akt signalling pathway	2.57E-27
Th17 cell differentiation	8.67E-20
MAPK signalling pathway	1.33E-18
Human T-cell leukemia virus 1 infection	4.6E-18
Osteoclast differentiation	4.73E-17
AGE-RAGE signalling pathway in diabetic complications	5.56E-17
Rap1 signalling pathway	6.65E-16
PD-L1 expression and PD-1 checkpoint pathway in cancer	1.06E-15

**Table 11.** Top Enriched KEGG Pathways in Lung in Symmetric Differential Network.





**Fig. 15.** DCNs M-F(1): for a better visualization, we represent only a fragment of the networks. The whole networks are available as supplementary material. **(a)** DCN M-F in amigdala tissue. **(b)** DCN M-F in aorta tissue. **(c)** DCN M-F in coronary tissue. **(d)** DCN M-F in adipose tissue.

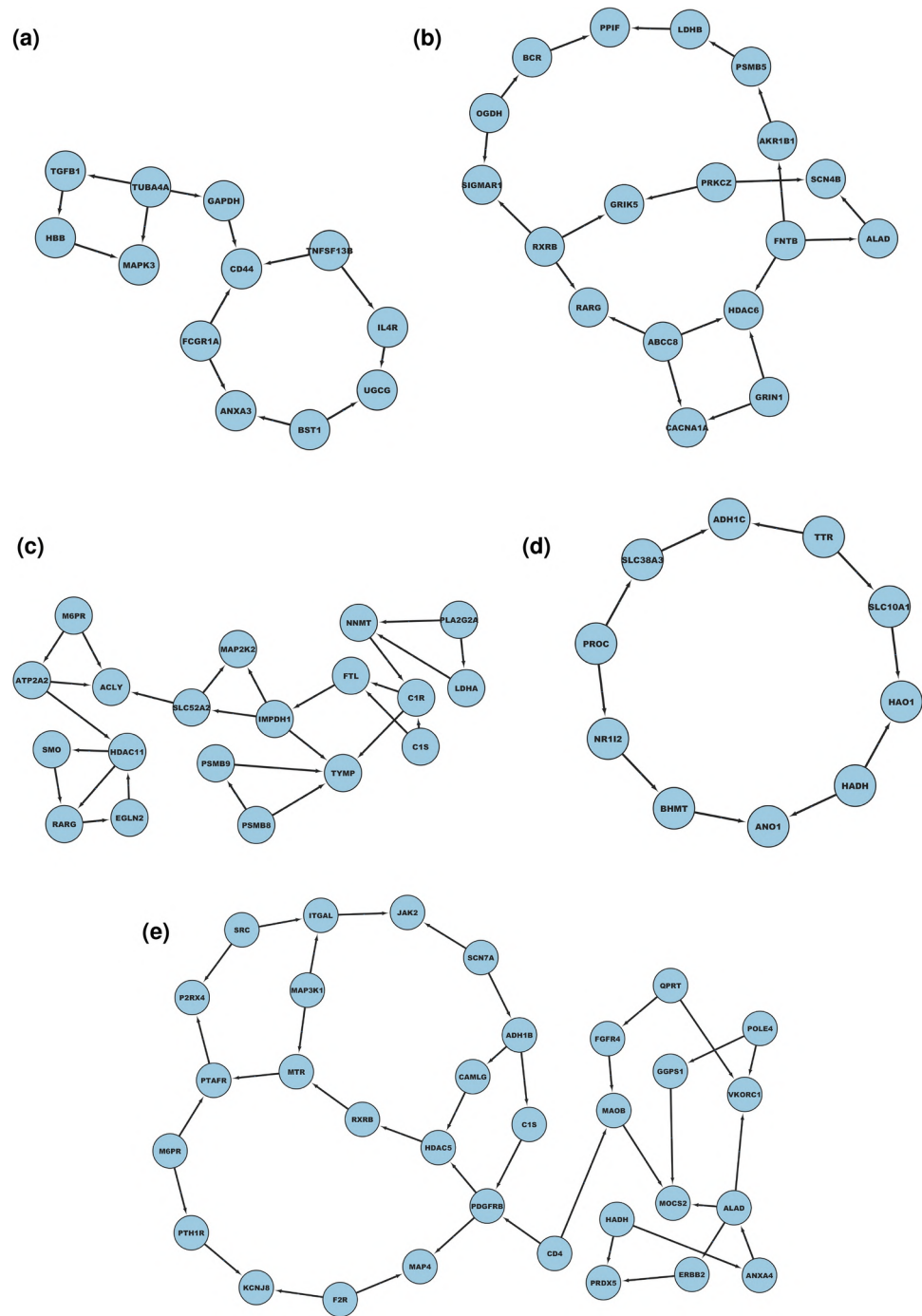
include the PI3K-Akt signalling pathway, MAPK signalling pathway, implicating their roles in adipocyte function, insulin sensitivity, and obesity-related complications in diabetic patients.

### Differential Networks obtained as Female Minus Male (F-M)

In this section, the differential networks obtained as Female minus Males are illustrated in the Figures 17, 18, 19 for the selected tissues. As before, for each network we found the enriched pathways as shown in Tables 21,22,23,24,25,26,27,28,29.

### Baseline comparison

To show the robustness of the approach, we compared the obtained results (male vs female) to a baseline obtained by random sampling data from the experimental dataset to show the differences between the networks. As a proof of principle, the robustness of the built DCN is related to the robustness of the methods used for creating the causal networks corresponding to the compared system. Consequently, the robustness of the results



**Fig. 16.** DCNs M-F(2): for a better visualization, we represent only a fragment of the networks. The whole networks are available as supplementary material. **(a)** DCN M-F in blood tissue. **(b)** DCN M-F in cerebellum tissue. **(c)** DCN M-F in colon tissue. **(d)** DCN M-F in liver tissue. **(e)** DCN M-F in lung tissue.

is somewhat independent from the number of the edges of the DCN. A DCN with zero edges DCN means that the compared systems have no difference.

Nevertheless, we performed a baseline experiment in which we sampled patients at random, independently of their sex. We therefore obtained a DCN for each pair of networks obtained from random sampling. The following algorithm was applied: given a dataset  $D$ , we consider  $N=100$  pairs of random sets of patients. We generate the DCNs for each pair and measure the number of edges. Finally, we compared the results using a non-parametric test (Wilcoxon's test) to assess whether significant differences exist in the number of edges obtained with a random split concerning the number of edges of the DCNs obtained with the original split (male/female). Taking a random tissue from among those analysed, we conducted the analysis just described considering 100 tests. We measured the number of edges for each random DCN and showed that the number

Pathway	FDR value
PI3K-Akt signalling pathway	1.49E-31
MAPK signalling pathway	1.97E-20
cGMP-PKG signalling pathway	1.41E-19
HIF-1 signalling pathway	7.05E-19
Rap1 signalling pathway	1.95E-18
AGE-RAGE signalling pathway in diabetic complications	2.05E-18
Proteoglycans in cancer	5.04E-18
Osteoclast differentiation	1.06E-17
EGFR tyrosine kinase inhibitor resistance	6.61E-17

**Table 12.** Top Enriched KEGG Pathways in Adipose Tissue in M-F Differential Network.

Pathway	FDR value
Alzheimer disease	2.79E-13
MAPK signalling pathway	7.72E-13
Oxytocin signalling pathway	2.39E-11
cAMP signalling pathway	5.44E-10
Thyroid hormone signalling pathway	1.49E-09
Adrenergic signalling in cardiomyocytes	1.72E-09
Rap1 signalling pathway	1.87E-09
Gap junction	3.36E-09

**Table 13.** Top Enriched KEGG Pathways in amigdala in M-F Differential Network.

Pathway	FDR value
PI3K-Akt signalling pathway	4.35E-19
Osteoclast differentiation	3.92E-13
Proteoglycans in cancer	3.97E-13
MAPK signalling pathway	5.67E-13
cGMP-PKG signalling pathway	6.95E-13
AGE-RAGE signalling pathway in diabetic complications	6.95E-13
Th17 cell differentiation	9.32E-13
HIF-1 signalling pathway	1.39E-12

**Table 14.** Top Enriched KEGG Pathways in Aorta in M-F Differential Network.

Pathway	FDR value
Th17 cell differentiation	5.72E-24
Osteoclast differentiation	2.6E-24
PD-L1 expression and PD-1 checkpoint pathway in cancer	8.06E-22
Hematopoietic cell lineage	2.49E-20
Human T-cell leukemia virus 1 infection	2.49E-20
PI3K-Akt signalling pathway	1.28E-17
Leishmaniasis	1.69E-16
Th1 and Th2 cell differentiation	3.65E-16
Tuberculosis	6.67E-16

**Table 15.** Top Enriched KEGG Pathways in Blood in M-F Differential Network.

of edges is significantly lower (p-value less than 0.05). Table 30 indicates the number of edges in the liver tissue, and Table 31 the corresponding p-value.

We also show the qualitative comparison of the DCN corresponding to male-female in relation to the 100 random splits. For each DCN, we take the Jaccard Index of the set of edges, and we report in Figure 20 the

Pathway	FDR value
Oxytocin signalling pathway	2.04E-17
MAPK signalling pathway	4.44E-16
Alzheimer disease	1.15E-15
cGMP-PKG signalling pathway	1.34E-11
cAMP signalling pathway	1.34E-11
Prion disease	8.38E-10
Adrenergic signalling in cardiomyocytes	1.66E-09
GABAergic synapse	1.66E-09
Serotonergic synapse	2.38E-09

**Table 16.** Top Enriched KEGG Pathways in Cerebellum in M-F Differential Network.

Pathway	FDR value
PI3K-Akt signalling pathway	5.32E-17
cGMP-PKG signalling pathway	1.01E-15
Oxytocin signalling pathway	1.01E-15
MAPK signalling pathway	4.02E-15
MicroRNAs in cancer	4.31E-14
Alzheimer disease	1.01E-13
GFR tyrosine kinase inhibitor resistance	5.15E-13
Central carbon metabolism in cancer	1.26E-12

**Table 17.** Top Enriched KEGG Pathways in Colon in M-F Differential Network.

Pathway	FDR value
PI3K-Akt signalling pathway	1.27E-21
MAPK signalling pathway	6.93E-16
Osteoclast differentiation	1.39E-14
Proteoglycans in cancer	1.61E-14
AGE-RAGE signalling pathway in diabetic complications	2.43E-14
Th17 cell differentiation	3.67E-14
Kaposi sarcoma-associated herpesvirus infection	4E-15
Focal adhesion	9.08E-14

**Table 18.** Top Enriched KEGG Pathways in Coronary in M-F Differential Network.

Pathway	FDR value
Complement and coagulation cascades	3,5E-26
Carbon metabolism	0,00000282
Alzheimer disease	0,0000016
Tyrosine metabolism	0,0000309
Th17 cell differentiation	0,0000327
Fatty acid degradation	0,0000812
PI3K-Akt signaling pathway	0,000247
Non-alcoholic fatty liver disease	0,000917

**Table 19.** Top Enriched KEGG Pathways in Liver in M-F Differential Network.

median value of the index. The values show low similarity among the real DCN with respect to the random DCN.

These results provide further foundation to the robustness and validity of the approach used, indicating that sex-based division does indeed produce more informative and meaningful network structures than random division.



Pathway	FDR value
PI3K-Akt signalling pathway	1.89E-27
Th17 cell differentiation	7.34E-20
MAPK signalling pathway	1.06E-18
Human T-cell leukemia virus 1 infection	3.81E-17
Osteoclast differentiation	4.04E-17
AGE-RAGE signalling pathway in diabetic complications	4.8E-18
Rap1 signalling pathway	5.58E-16
PD-L1 expression and PD-1 checkpoint pathway in cancer	9.25E-16

**Table 20.** Top Enriched KEGG Pathways in Lung in M-F Differential Network.

## Discussion

This study presents the development and application of Differential Causal Networks (DCNs), which offer a unique approach to understanding the complex causal relationships in biological systems, particularly in the context of Type 2 Diabetes Mellitus (T2DM). DCNs allow for a more nuanced exploration of the dynamic and multifaceted nature of biological processes by focusing on the differences in causal relationships between two systems or conditions. This discussion explores the implications, challenges, and future directions of the study on DCNs.

The application of DCNs to investigate sex-specific differences in gene expression related to T2DM is a significant contribution to network medicine and personalized healthcare. By identifying differential causal relationships between male and female patients across various tissues, this study enhances our understanding of the pathophysiology of T2DM and underscores the importance of considering sex as a critical factor in disease research and treatment. The insights gained from DCN analysis could guide the development of more targeted therapeutic interventions, taking into account the unique causal mechanisms at play in different patient subgroups.

The methodological framework introduced for constructing and analyzing DCNs represents a significant advancement in causal inference and network analysis. However, the approach is not without its challenges. One primary concern is the need for large-scale, high-quality experimental data to accurately infer causal relationships. Additionally, the complexity of DCNs requires sophisticated computational tools and algorithms, posing a barrier to wider adoption without accessible software and clear guidelines for use. Future work should focus on refining the DCN construction process, improving computational efficiency, and developing more user-friendly tools for researchers across disciplines.

There are several promising directions for extending the application of DCNs beyond the study of T2DM. One potential area is the exploration of differential causal mechanisms across a broader range of diseases, including cancer, cardiovascular diseases, and neurodegenerative disorders. Furthermore, DCNs could be applied to investigate the impact of environmental factors, lifestyle choices, and other exogenous variables on disease progression and treatment outcomes. Another exciting avenue for future research is the integration of DCNs with other data types, such as proteomics and metabolomics data, to construct more comprehensive models of disease mechanisms.

## Biological relevance of differential networks

### *Biological Relevance of M-F*

The tables 12, 13, 14, 15, 16, 17, 18, 19, 20 present a comprehensive analysis of KEGG pathway enrichment in different tissues within a male-female (M-F) differential network. This analysis illuminates significant pathways that may underlie sex-specific differences in disease predisposition, progression, and response to treatments, particularly focusing on diabetes comorbidities and cancer. Below is a summary and discussion of the findings from each tissue analyzed.

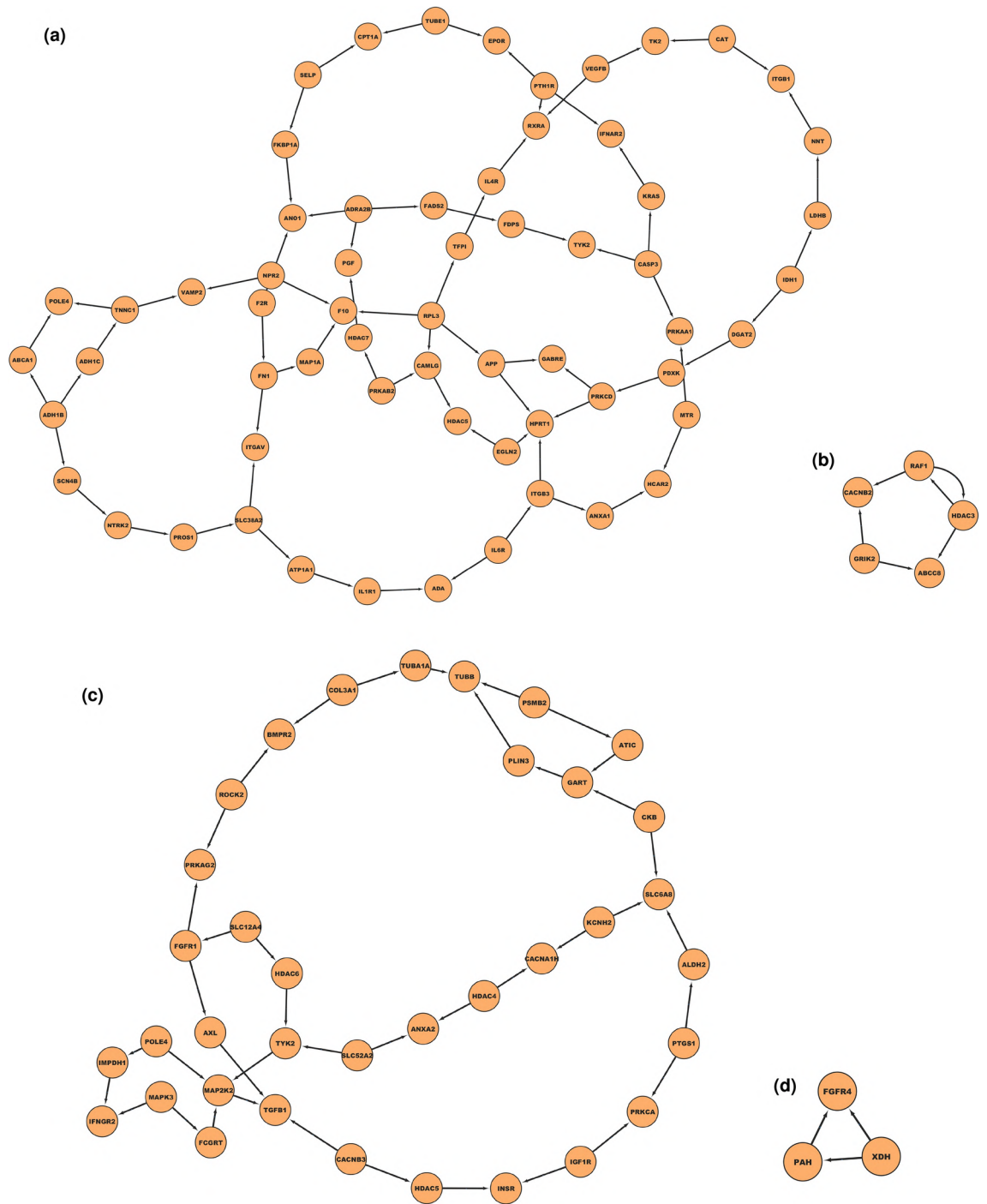
Adipose tissue significantly enriched pathways such as the PI3K-Akt signalling pathway, MAPK signalling pathway, and AGE-RAGE signalling pathway in diabetic complications. The prominence of the PI3K-Akt and MAPK signalling pathways suggests a crucial role in cell survival, proliferation, and metabolism, which are vital in adipose tissue function and dysfunction. The AGE-RAGE signalling pathway's enrichment highlights the tissue's involvement in diabetes complications, reflecting the metabolic and inflammatory processes diabetes affects in adipose tissue.

Significant pathways in the amygdala include those associated with Alzheimer's disease, the MAPK signalling pathway, and the oxytocin signalling pathway. The enrichment of Alzheimer's disease pathways underscores the amygdala's role in neurodegenerative processes. Oxytocin signalling suggests its importance in emotional and stress responses, which are pertinent to the amygdala's function in the brain. The presence of neuroinflammatory and pathways differentially structured in our causal networks may reflect the observed sex-related vulnerabilities to neurodegenerative diseases<sup>43</sup>.

The aorta showed enrichment in pathways such as PI3K-Akt signalling and AGE-RAGE signalling in diabetic complications, suggesting a complex interplay of proliferative and metabolic pathways contributing to vascular health and disease. This underlines the vascular system's vulnerability to diabetic complications.

Enrichment in Th17 cell differentiation within the blood highlights the immune system's role in mediating inflammation and its potential dysregulation in diseases. In addition, Osteoclast differentiation is responsible for



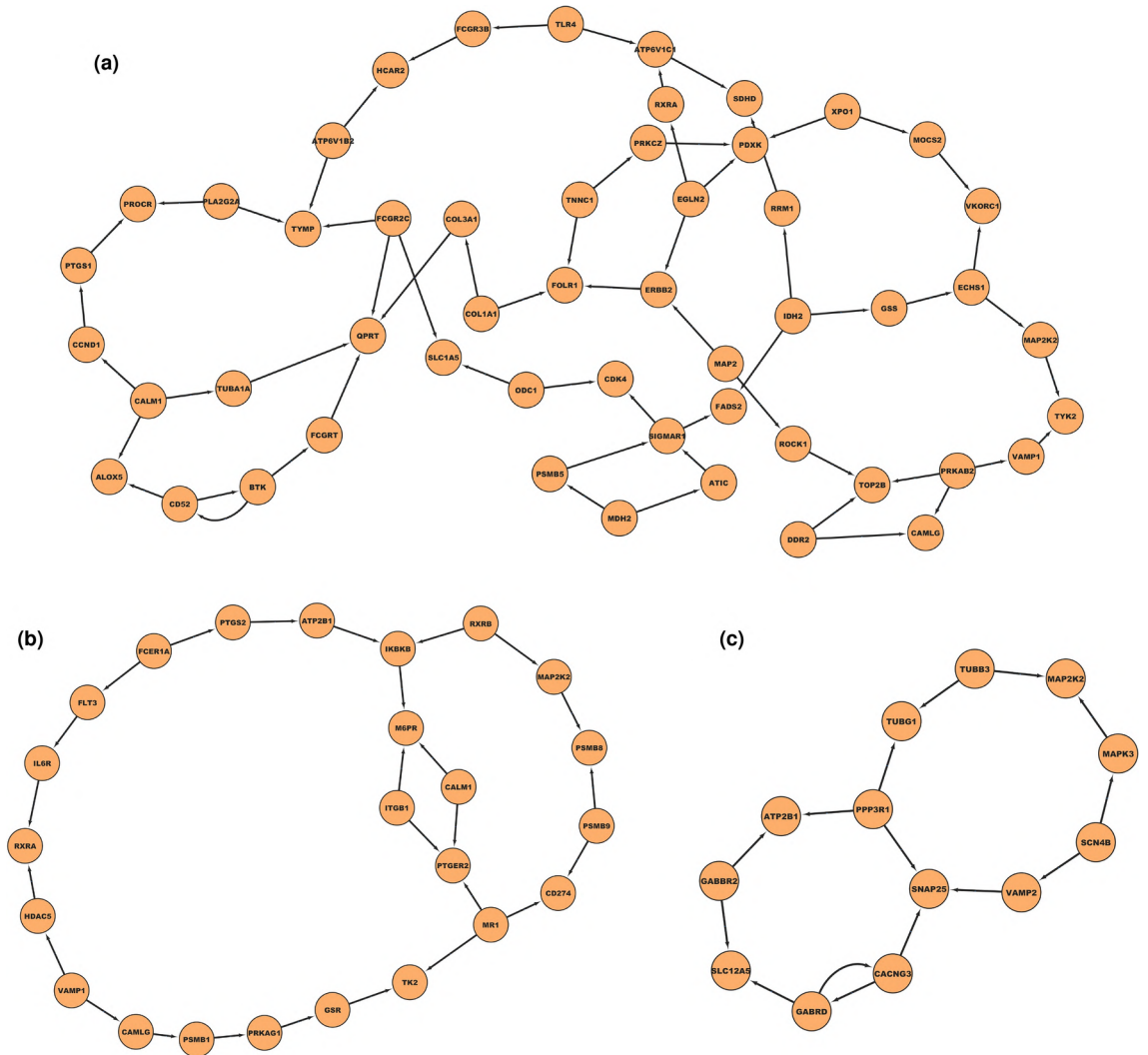


**Fig. 18.** DCNs F-M(2): for a better visualization, we represent only a fragment of the networks. The whole networks are available as supplementary material. **(a)** DCN F-M in adipose tissue. **(b)** DCN F-M in cerebellum tissue. **(c)** DCN F-M in colon tissue. **(d)** DCN F-M in liver tissue.

PI3K-Akt signalling and cGMP-PKG signalling pathways were notably enriched in the colon<sup>44</sup>. This underscores the importance of these signalling pathways in maintaining intestinal barrier function and responding to microbial and inflammatory stimuli.

The coronary artery analysis revealed enrichment in PI3K-Akt signalling, MAPK signalling pathway and osteoclast differentiation. These findings reflect the coronary artery’s critical role in cardiovascular disease, with implications for atherosclerosis and coronary artery disease development.

The liver’s enriched pathways cover various metabolic processes, reflecting its central role in metabolism, detoxification, and systemic regulation. Notable pathways include carbon metabolism



**Fig. 19.** DCNs F-M(3): for a better visualization, we represent only a fragment of the networks. The whole networks are available as supplementary material. (a) DCN F-M in lung tissue. (b) DCN F-M in blood tissue. (c) DCN F-M in amigdala tissue.

Pathway	FDR value
PI3K-Akt signalling pathway	1.49E-29
MAPK signalling pathway	1.97E-18
cGMP-PKG signalling pathway	1.41E-17
HIF-1 signalling pathway	7.05E-17
Rap1 signalling pathway	1.95E-16
AGE-RAGE signalling pathway in diabetic complications	2.05E-16
Proteoglycans in cancer	5.04E-16
Osteoclast differentiation	1.06E-15

**Table 21.** Top Enriched KEGG Pathways in Adipose in F-M Differential Network.

emphasizing the liver’s critical functions and its susceptibility to metabolic diseases. PI3K-Akt signalling and Th17 cell differentiation were significantly enriched in the lung. These findings highlight the lung’s vulnerability to cancer and the role of immune signalling in respiratory diseases, which may differ between sexes.

Across tissues, common pathways such as the PI3K-Akt signalling pathway and MAPK signalling pathway, were frequently enriched, suggesting these pathways’ fundamental roles in mediating sex-specific differences in disease mechanisms. The repeated identification of signalling pathways related to cell growth, survival, and

Pathway	FDR value
Alzheimer disease	2.79E-13
Oxytocin signalling pathway	3.51E-12
MAPK signalling pathway	5.02E-12
Thyroid hormone signalling pathway	1.5E-11
cAMP signalling pathway	4.53E-10
Adrenergic signalling in cardiomyocytes	1.72E-09
Gap junction	3.73E-09
cGMP-PKG signalling pathway	7.01E-09

**Table 22.** Top Enriched KEGG Pathways in amigdala in F-M Differential Network.

Pathway	FDR value
PI3K-Akt signaling pathway	4,99E-20
Proteoglycans in cancer	5,14E-14
MAPK signaling pathway	8,7E-15
cGMP-PKG signaling pathway	8,7E-15
Osteoclast differentiation	2,24E-13
FoxO signaling pathway	4,87E-13
AGE-RAGE signaling pathway in diabetic complications	5,4E-14
Influenza A	5,4E-14

**Table 23.** Top Enriched KEGG Pathways in Aorta in F-M Differential Network.

Pathway	FDR value
Th17 cell differentiation	9.22E-24
Osteoclast differentiation	2.33E-23
PD-L1 expression and PD-1 checkpoint pathway in cancer	7.32E-22
Human T-cell leukemia virus 1 infection	2.6E-21
Hematopoietic cell lineage	3.32E-19
PI3K-Akt signalling pathway	9E-18
Leishmaniasis	1.57E-16
Th1 and Th2 cell differentiation	3.38E-16
Tuberculosis	6.09E-16

**Table 24.** Top Enriched KEGG Pathways in Blood in F-M Differential Network.

Pathway	FDR value
Oxytocin signalling pathway	2.73E-16
Alzheimer disease	1.49E-15
MAPK signalling pathway	2.58E-15
cGMP-PKG signalling pathway	1.52E-11
cAMP signalling pathway	9.61E-11
Prion disease	8.9E-11
Adrenergic signalling in cardiomyocytes	1.78E-09
GABAergic synapse	1.78E-09
Serotonergic synapse	2.23E-09

**Table 25.** Top Enriched KEGG Pathways in Cerebellum in F-M Differential Network.



Pathway	FDR value
PI3K-Akt signalling pathway	5.32E-17
cGMP-PKG signalling pathway	1.01E-15
Oxytocin signalling pathway	1.01E-15
MAPK signalling pathway	4.02E-15
MicroRNAs in cancer	4.31E-14
Alzheimer disease	1.01E-13
EGFR tyrosine kinase inhibitor resistance	5.15E-13
Central carbon metabolism in cancer	1.26E-12

**Table 26.** Top Enriched KEGG Pathways in Colon in F-M Differential Network.

Pathway	FDR value
PI3K-Akt signalling pathway	1.33E-22
MAPK signalling pathway	7.74E-17
Osteoclast differentiation	1.39E-14
Proteoglycans in cancer	1.61E-14
FoxO signalling pathway	2.4E-15
AGE-RAGE signalling pathway in diabetic complications	2.43E-14
cGMP-PKG signalling pathway	2.44E-14
Th17 cell differentiation	3.27E-14

**Table 27.** Top Enriched KEGG Pathways in Coronary in F-M Differential Network.

Pathway	FDR value
Complement and coagulation cascades	3.5E-26
Carbon metabolism	0.00000282
Alzheimer disease	0.0000016
Tyrosine metabolism	0.0000309
Th17 cell differentiation	0.0000327
Fatty acid degradation	0.0000812
PI3K-Akt signalling pathway	0.000247
Non-alcoholic fatty liver disease	0.000917

**Table 28.** Top Enriched KEGG Pathways in Liver in F-M Differential Network.

Pathway	FDR value
PI3K-Akt signalling pathway	1.89E-27
Th17 cell differentiation	1.08E-18
MAPK signalling pathway	8.76E-18
Human T-cell leukemia virus 1 infection	3.5E-17
Osteoclast differentiation	4.69E-16
Rap1 signalling pathway	5.74E-16
AGE-RAGE signalling pathway in diabetic complications	5.74E-16
PD-L1 expression and PD-1 checkpoint pathway in cancer	1.27E-14

**Table 29.** Top Enriched KEGG Pathways in Lung in F-M Differential Network.

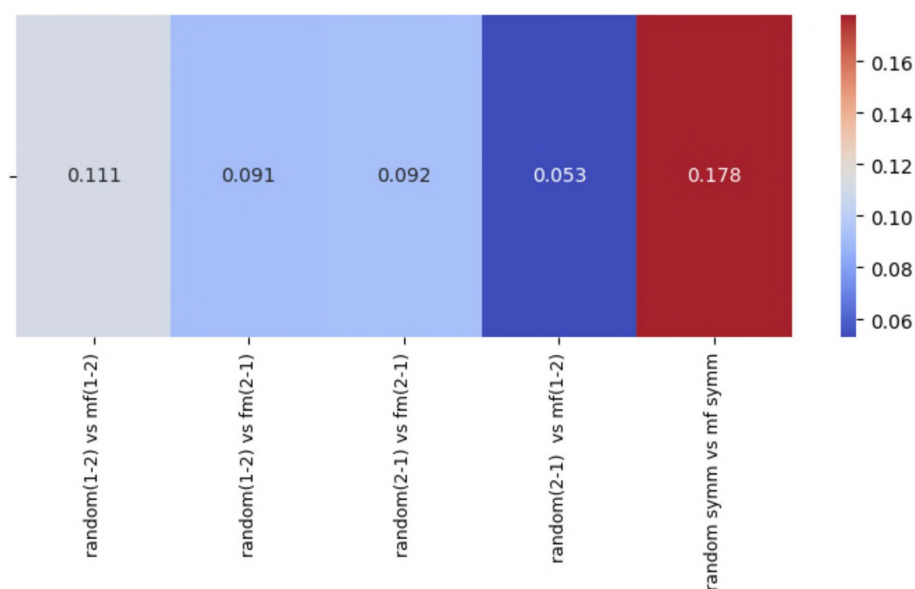
inflammation underscores their significance in various tissues and diseases, including diabetes and cancer. The AGE-RAGE signalling pathway's prominence in diabetic complications across tissues highlights the systemic impact of diabetes on different organs. The analysis showcases the intricate relationship between sex-specific molecular mechanisms and disease, emphasizing the need for tailored approaches to understanding and treating diseases in men and women.

	Symmetric DCNs	DCNs Group1 - Group2	DCNs Group2 - Group1
Number of Edges in the Male-Female Network	554	337	217
Median Number of Edges in Random Networks	530	270	170

**Table 30.** Comparison of the Number of the Edges in the male-female group with respect to random splits.

	Symmetric DCNs	DCNs Group1 - Group2	DCNs Group2 - Group1
p-value	5.2e-3	3.81e-18	3.83e-18

**Table 31.** Statistical significance of observed differences between male-female split and random split for Wilcoxon's test.



**Fig. 20.** Median Values of the Jaccard indexes were obtained by comparing random differential networks with respect to the male-female network. The low values of these indexes represent that the real network shares a low number of edges concerning the random ones. Given the networks obtained with male-female splits (i.e. male vs females, females vs males and symmetric) and the random networks (1 vs 2, 2 vs 1 and symmetric), we measure all the possible Jaccard indexes, and we report the median values of each group.

#### Biological relevance of F-M

The analysis of KEGG pathways in different tissues within a F-M differential network sheds light on the molecular underpinnings of sex-specific differences in various biological processes and disease susceptibilities. Here, we discuss the top enriched pathways identified in adipose, amygdala, blood, cerebellum, colon, coronary, liver, and lung tissues and reported in the Tables 21, 22, 23, 24, 25, 26, 27, 28, 29.

Enrichment in pathways like the PI3K-Akt signalling and MAPK signalling underscores their critical role in adipose tissue's metabolic and cell signalling processes, influencing obesity, insulin resistance, and diabetes mellitus. The significance of the AGE-RAGE signalling pathway points to the tissue's involvement in inflammatory responses and diabetic complications.

The enrichment of the oxytocin signalling pathway, alongside Alzheimer's disease and MAPK signalling pathways, highlights the amygdala's involvement in emotional processing, stress responses, and neurodegenerative diseases. These pathways could explain sex differences in emotional behaviour and susceptibility to mental health disorders.

PI3K-Akt signalling pathway and Proteoglycans in cancer in aorta tissue promote cancer growth and progression by regulating cellular signalling and other interactions.

Significant pathways like Th17 cell differentiation and PD-L1 expression indicate the blood's crucial role in mediating immune responses and inflammation, which can vary between sexes and influence disease outcomes. Cancer-related pathways also suggest systemic effects on tumour progression and immune evasion.

The enrichment of the oxytocin and MAPK signalling pathways suggests their importance in cerebellar functions related to motor control and cognitive processes. Alzheimer's disease pathway enrichment further implicates sex differences in neurodegenerative disease progression and susceptibility.

Cancer-related pathways, PI3K-Akt signalling, and cGMP-PKG signalling pathway enrichment in the colon underline the tissue's role in colorectal cancer and highlight potential mechanisms behind sex-specific differences in cancer incidence and progression.

In the coronary tissue, PI3K-Akt signalling and osteoclast differentiation are enriched, indicating mechanisms that may contribute to sex differences in cardiovascular diseases, including atherosclerosis and coronary artery disease.

The enrichment of carbon metabolism in the liver underscores the crucial role of this complex set of biochemical processes for maintaining metabolic homeostasis, regulating blood glucose concentration, and performing detoxification functions.

The lung shows a high enrichment of PI3K-Akt signalling, suggesting critical roles in lung physiology and pathology, including cancer. The Th17 cell differentiation pathway's significance points to sex differences in immune responses, potentially affecting susceptibility to respiratory infections and diseases.

Common themes across tissues include the significant enrichment of signalling (e.g., PI3K-Akt, MAPK) and immune response, underscoring fundamental sex-specific molecular differences. These findings highlight the importance of considering sex as a biological variable in biomedical research. They may provide insights into tailored therapeutic strategies for treating various diseases in men and women.

Research on DCNs and other network-based approaches aims to inform clinical practice and improve patient outcomes. To achieve this, bridging the gap between theoretical models and practical applications is essential. This involves validating the findings from DCN analysis in clinical studies, exploring the potential for DCNs to predict disease progression and response to treatment, and developing guidelines for incorporating DCN insights into diagnostic and therapeutic decision-making processes.

#### *Biological relevance of sym differential networks*

The tables 3, 4, 5, 6, 7, 8, 9, 10, 11 delineate the top enriched Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways in various tissues within a symmetric differential network, highlighting crucial biological processes and signalling pathways that may underpin physiological and pathological differences at the cellular and molecular levels. These findings could offer insights into the mechanisms behind diseases, tissue functions, and potential therapeutic targets. Let's discuss these results.

The adipose tissue shows significant enrichment in pathways such as the PI3K-Akt and MAPK signalling pathways, central to cell growth, survival, and metabolism, suggesting their pivotal roles in adipose function, obesity, insulin resistance, and inflammation. The AGE-RAGE signalling pathway's enrichment indicates its involvement in diabetic complications, highlighting the adipose tissue's contribution to metabolic diseases.

Enrichments in the amygdala include Alzheimer's disease and the oxytocin signalling pathway, pointing to the amygdala's involvement in neurodegenerative processes and emotional and social behavior regulation. The presence of cancer pathways and thyroid hormone signalling suggests a broader spectrum of physiological roles and potential vulnerabilities of the amygdala to various diseases.

PI3K-Akt signalling, and proteoglycans in cancer are notably enriched in the aorta, emphasizing the vascular system's susceptibility to oncogenic processes and signalling that may influence vascular health, atherosclerosis, and diabetes-related vascular complications.

Significant pathways in the blood include those related to immune response (PD-L1 expression and PD-1 checkpoint pathway), and infectious diseases (Human T-cell leukemia virus 1 infection), reflecting the blood's role in immune surveillance, inflammation, and systemic responses to pathogens and malignancies.

The cerebellum shows enrichment in the oxytocin signalling pathway, MAPK signalling pathway, and Alzheimer's disease, underscoring its involvement in cognitive functions, neuroprotection, and emotional processing. This may indicate sex-specific differences in cerebellar functions and disease susceptibilities.

The colon's enriched pathways highlight the PI3K-Akt and MAPK signalling pathways is crucial for cell proliferation and survival. The presence of pathways related to microbial infections (e.g., microRNAs in cancer) suggests interactions between the microbiome and colon cancer development.

In the coronary tissue, pathways related to signalling (PI3K-Akt, MAPK), underscore the heart's exposure to signalling alterations in cardiovascular diseases affecting heart function.

The liver's enriched pathways cover a wide range of metabolic processes, reflecting its central role in metabolism, detoxification, and systemic regulation. Notable pathways include carbon metabolism, emphasizing the liver's critical functions and its susceptibility to metabolic diseases.

Enriched pathways in the lung, such as those involved in PI3K-Akt signalling, and Th17 cell differentiation, highlight the lung's involvement in immune responses, inflammation, and susceptibility to cancer. This points to potential mechanisms driving lung diseases and the impact of environmental factors on lung health.

## Methods

The T2DiACoD database was utilized to extract a compilation of genes associated with diabetes comorbidities, as detailed in<sup>45</sup>. Gene expression data were procured from the GTEx database<sup>29</sup>.

T2DiACoD is a meticulously curated repository housing genes and non-coding RNAs linked to Type 2 Diabetes Mellitus (T2DM). Its contents are amassed through manual curation from literature and existing databases, with a particular focus on comorbidities such as atherosclerosis, nephropathy, diabetic retinopathy, and cardiovascular diseases. The database encompasses 650 genes and 34 microRNAs associated with these comorbidities.

The Genotype-Tissue Expression (GTEx) data portal serves as a publicly accessible repository for genomic data from various individuals. This resource includes diverse genomic data types, ranging from sequencing to methylation.

GTEX furnishes metadata for each sample, including details on tissue of origin, sex, and age (grouped into six categories). Consequently, GTEX serves as a crucial resource for studies examining age-tissue dependencies. As of February 1st, the current version of the GTEX database contains 17,382 samples across 54 tissues from 948 donors, accessible through the web interface provided by the GTEX portal. This interface facilitates seamless data querying and visualisation<sup>15,46,47</sup>. Additionally, data can be downloaded and analyzed using custom scripts.

Our study used data from various tissues, creating two classes stratified by sex. We used TPM expression data by tissue of GTEX v8 (GTEX v8 TPM expression data) We analyzed the pharmacologically active genes implicated in the onset of complications related to type 2 diabetes mellitus across 9 tissues (blood tissue, brain tissue, adipose tissue, amygdala tissue, aorta tissue, colon tissue, coronary tissue, liver tissue, and lung tissue).

### Data preprocessing

We preprocessed the raw GTEX data by eliminating genes presenting expression close to zero in a significant number of samples. In particular, we eliminated genes that have expression in the range [0-20] in at least a stated number of samples ( $n = 400$  for adipose tissue;  $n = 100$  for amygdala, coronary, liver, and cerebellum tissue;  $n = 200$  for aorta and colon tissue;  $n = 500$  for blood tissue;  $n = 300$  for lung tissue), obtaining, for each tissue, the number of genes specified in the table for the analysis. Having done that, we obtained the data in CSV format, upon which the causality analysis described in the following section was conducted. We split the latter into two datasets according to sex, obtaining, in numerical terms, the data summarized in Table 32. It is important to note that the number of female samples is lower than that of male samples for all the tissues analyzed so that this factor could influence the analysis.

### Causal discovery on T2DM dataset

For each tissue, we applied the PC algorithm<sup>30</sup> on gene expression data to infer causal relationships separately between males and females. The PC algorithm (Peter-Clark algorithm) identifies the causal structure of a set of variables from observational data by analysing a series of conditional independence tests to determine the relationships between variables, ultimately constructing a directed network that represents these causal relationships. In our experiments, we set the default parameters of PC.

After, we calculated a DCN for each pair of corresponding networks. We used the gCastle package<sup>48</sup> which provides different algorithms for PC. We selected the algorithm proposed by Kalisch et al., since it presents optimal performances as described in literature<sup>49</sup>. We used default parameters: variant = 'original', ci test = 'fisherz' (independent test default) and no background information added (priori knowledge: None). In addition, we set  $\alpha = 0.05$  (significance level default) according to M.Kalisch and P.Buhlmann's tests<sup>49</sup>.

Experimental tests have been performed on a 8 CPUs Intel Xeon W5, 32GB DDR5 RAM. We used the *DiGraph* class of *NetworkX* library for building and visualising causal networks and for calculating the differential causal networks using respectively *difference* and *symmetric difference* functions between male and female networks. For each network, we analysed the pathway enrichment considering the KEGG<sup>38</sup> (Kanehisa 2002) pathway database and Biological Process of Gene Ontology<sup>50</sup> by using the STRING enrichment app of the Cytoscape software<sup>41</sup>. Data and Code are available at <https://github.com/hguzzi/DifferentialCausalNetworks.git>.

### Conclusion

In this work, we introduce DCNs, to model causal differences and to capture known and unknown relationships among genes between sexes. This innovative method promises to shed new light on the molecular basis of Type 2 Diabetes Mellitus. We extracted genes from nine tissues associated with Type 2 Diabetes Mellitus comorbidities from T2DiACoD database and the relative expression in male and female individuals from GTEX. After data preprocessing and partitioning by sex, we applied the PC algorithm as a causal discovery method, resulting in two networks for each tissue (one for males and one for females). Subsequently, we derived DCNs by performing three types of differences between male and female networks. Consequentially, we obtained three DCNs: male minus female differential causal networks, female minus male differential causal networks and symmetrical differential causal networks. As explained, the results reveal that the calculation of a DCN is not symmetric performing simple differences. Finally, we conducted a comprehensive pathway enrichment analysis for each

Tissue	Total n. of samples	Males samples	Females samples	n.genes analysed
Adipose tissue	451	311	140	313
amigdala tissue	118	86	32	195
Aorta tissue	281	181	100	279
Blood tissue	512	346	166	283
Cerebellum tissue	193	145	48	225
Colon tissue	295	195	100	251
Coronary tissue	153	91	62	269
Liver tissue	158	117	41	187
Lung tissue	385	273	112	329

**Table 32.** Resuming of the data obtained after the preprocessing phase where both numbers of analyzed samples and genes are reported.

DCN, utilizing the KEGG pathway database and Biological Process of Gene Ontology, using the Cytoscape software in conjunction with its STRING enrichment app.

The study delves into Differential Causal Networks (DCNs) as a new method to comprehend the basis of Type 2 Diabetes Mellitus (T2DM) across various tissues, focusing on the differences between males and females. Combining data from the T2DiACoD and GTEx databases, the research carefully constructs causal networks to highlight gene expression disparities between males and females. This extensive analysis uncovers the diverse molecular pathways implicated in T2DM and offers insights into its complex biological mechanisms.

Using DCNs to investigate the causal underpinnings in different tissues uncovers significant pathways such as PI3K-Akt and MAPK signalling, emphasizing their critical roles in cell metabolism, growth, and survival. The study emphasizes pathways like AGE-RAGE in diabetic complications, highlighting the multifaceted impact of diabetes on different organs. The analysis also goes beyond metabolic pathways, exploring the involvement of Alzheimer's disease pathways and the oxytocin signalling pathway, suggesting a broader spectrum of biological processes influenced by sex differences in T2DM.

Enrichment analysis across tissues provides critical insights into different organs' physiological and pathological roles in T2DM. For instance, the study highlights the role of adipose tissues in metabolic dysregulation and the susceptibility of the aorta to vascular complications. Similarly, the analysis highlights the role of processes within the blood in immune mediation, as well as the central role of the genes expressed in liver for regulating metabolism. The findings also draw attention towards potential links between T2DM, sex differences, and cancer risk. The here presented approach based on generating and analyzing DCNs, offers new opportunities for comprehending the complexity of T2DM.

The primary aim of researching DCNs and other network-based methodologies in a biological context is to enhance clinical care. Given this, our research opens the doors to using DCNs in the genetic field to personalize the care of patients suffering from pathologies such as type II diabetes mellitus analyzed in this study. Specifically, they support the differentiation between conditions, such as the two sexes, with the benefit of examining the causal mechanisms that influence certain pathological conditions. With similar objectives, DCNs could be expanded into the proteomic and metabolomic domains. The clinical validation of the information drawn from this type of analysis remains of fundamental importance.

## Data availability

Code, Data and Supplementary Information are available at <https://github.com/hguzzi/DifferentialCausalNetworks>.

Received: 6 April 2024; Accepted: 2 November 2024

Published online: 09 November 2024

## References

- Barabási, A.-L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nature reviews genetics* **12**, 56–68 (2011).
- Zitnik, M. et al. Current and future directions in network biology. *Bioinformatics advances* **4**, vbae099. <https://doi.org/10.1093/bioadv/vbae099> (2024).
- Parini, P. et al. The network medicine imperative and the need for an international network medicine consortium. *The American journal of medicine* **133**, e451–e454 (2020).
- Gu, S., Jiang, M., Guzzi, P. H. & Milenković, T. Modeling multi-scale data via a network of networks. *Bioinformatics* **38**, 2544–2553 (2022).
- Cannataro, M., Guzzi, P. H. & Veltri, P. Protein-to-protein interactions: Technologies, databases, and algorithms. *ACM Computing Surveys (CSUR)* **43**, 1–36 (2010).
- Karlebach, G. & Shamir, R. Modelling and analysis of gene regulatory networks. *Nature reviews Molecular cell biology* **9**, 770–780 (2008).
- Cowen, L., Ideker, T., Raphael, B. J. & Sharan, R. Network propagation: a universal amplifier of genetic associations. *Nature Reviews Genetics* **18**, 551–562 (2017).
- Yang, Y. et al. Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nature communications* **5**, 3231 (2014).
- Lee, W.-P. & Tzou, W.-S. Computational methods for discovering gene networks from expression data. *Briefings in bioinformatics* **10**, 408–423 (2009).
- Lichtblau, Y. et al. Comparative assessment of differential network analysis methods. *Briefings in bioinformatics* **18**, 837–850 (2017).
- Ideker, T. & Krogan, N. J. Differential network biology. *Molecular systems biology* **8**, 565 (2012).
- Succurro, E. et al. Sex-specific differences in prevalence of nonalcoholic fatty liver disease in subjects with prediabetes and type 2 diabetes. *Diabetes Research and Clinical Practice* **190**, 110027 (2022).
- Guzzi, P. H. et al. Analysis of age-dependent gene-expression in human tissues for studying diabetes comorbidities. *Scientific Reports* **13**, 10372 (2023).
- Cannistraci, C. V., Valsecchi, M. G. & Capua, I. Age-sex population adjusted analysis of disease severity in epidemics as a tool to devise public health policies for covid-19. *Scientific reports* **11**, 1–8 (2021).
- Mercatelli, D., Pedace, E., Veltri, P., Giorgi, F. M. & Guzzi, P. H. Exploiting the molecular basis of age and gender differences in outcomes of sars-cov-2 infections. *Computational and Structural Biotechnology Journal* **19**, 4092–4100 (2021).
- Silverman, E. K. et al. Molecular networks in network medicine: Development and applications. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* **12**, e1489 (2020).
- Bang, S., Kim, J.-H. & Shin, H. Causality modeling for directed disease network. *Bioinformatics* **32**, i437–i444 (2016).
- Pearl, J. Causal inference in statistics: An overview. *Statist. Surv.* **3**, 96–146 (2009).
- Pearl, J. Causal inference without counterfactuals: Comment. *Journal of the American Statistical Association* **95**, 428–431 (2000).
- Pio, G., Ceci, M., Prisciandaro, F. & Malerba, D. Exploiting causality in gene network reconstruction based on graph embedding. *Machine Learning* **109**, 1231–1279 (2020).
- Schieber, T. A. et al. Quantification of network structural dissimilarities. *Nature communications* **8**, 13928 (2017).
- Ran, Y., Xu, X.-K. & Jia, T. The maximum capability of a topological feature in link prediction. *PNAS nexus* **3**, pgae113 (2024).



23. Jiang, Y., Li, M., Fan, Y. & Di, Z. Characterizing dissimilarity of weighted networks. *Scientific Reports* **11**, 5768 (2021).
24. Clark, M. Post congress tristesse. In *TeX90 Conference Proceedings*, 84–89 (TeX Users Group, 1991).
25. Guo, R., Cheng, L., Li, J., Hahn, P. R. & Liu, H. A survey of learning causality with data: Problems and methods. *ACM Computing Surveys (CSUR)* **53**, 1–37 (2020).
26. Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**, 688 (1974).
27. Succurro, E. et al. Impaired insulin-stimulated myocardial glucose metabolic rate is associated with reduced estimated myocardial energetic efficiency in subjects with different degrees of glucose tolerance. *Cardiovascular Diabetology* **22**, 1–11 (2023).
28. Guzzi, P. H. et al. Differential network analysis between sex of the genes related to comorbidities of type 2 mellitus diabetes. *Applied Network Science* **8**, 1–16 (2023).
29. Lonsdale, J. et al. The genotype-tissue expression (gtex) project. *Nature genetics* **45**, 580–585 (2013).
30. Glymour, C., Zhang, K. & Spirtes, P. Review of causal discovery methods based on graphical models. *Frontiers in Genetics* **10**, <https://doi.org/10.3389/fgene.2019.00524> (2019).
31. Molak, A. *Causal Inference and Discovery in Python: Unlock the secrets of modern causal machine learning with DoWhy, EconML, PyTorch and more* (Packt Publishing, Birmingham, 2023), 1. edn. <https://amzn.to/3RebWzn>.
32. Spirtes, P., Glymour, C. & Scheines, R. *Causation, Prediction, and Search*, vol. 81 (Adaptive Computation and Machine Learning/Thomas Dietterich, Editor; Christopher Bishop, David Heckerman, Michael Jordan, and Michael Kearns, Associate Editors, 1993).
33. Pearl, J. & Verma, T. Equivalence and synthesis of causal models. *Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence* 220–227 (1990).
34. Pearl, J. & Verma, T. A theory of inferred causation. In: Allen, J. A., Fikes, R., and Sand wall, E., editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference* 441–452 (1991).
35. Malinsky, D. & Danks, D. Causal discovery algorithms: A practical guide. *Philosophy Compass* **13**, e12470. <https://doi.org/10.1111/phc3.12470> (2018).
36. Foraita, R. et al. Causal Discovery of Gene Regulation with Incomplete Data. *Journal of the Royal Statistical Society Series A: Statistics in Society* **183**, 1747–1775. <https://doi.org/10.1111/rssa.12565> (2020). [https://academic.oup.com/jrssa/article-pdf/183/4/1747/49332541/jrssa\\_183\\_4\\_1747.pdf](https://academic.oup.com/jrssa/article-pdf/183/4/1747/49332541/jrssa_183_4_1747.pdf).
37. Kanehisa, M. & Goto, S. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research* **28**, 27–30 (2000).
38. Kanehisa, M. The kegg database. In *In Silico Simulation of Biological Processes: Novartis Foundation Symposium 247*, vol. 247, 91–103 (Wiley Online Library, 2002).
39. Kanehisa, M. Toward understanding the origin and evolution of cellular organisms. *Protein Science* **28**, 1947–1951 (2019).
40. Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M. & Ishiguro-Watanabe, M. Kegg for taxonomy-based analysis of pathways and genomes. *Nucleic acids research* **51**, D587–D592 (2023).
41. Doncheva, N. T., Morris, J. H., Gorodkin, J. & Jensen, L. J. Cytoscape stringapp: network analysis and visualization of proteomics data. *Journal of proteome research* **18**, 623–632 (2018).
42. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* **13**, 2498–2504 (2003).
43. Nebel, R. A. et al. Understanding the impact of sex and gender in alzheimer's disease: a call to action. *Alzheimer's & Dementia* **14**, 1171–1183 (2018).
44. Abenavoli, L. et al. Serum adipokine levels in overweight patients and their relationship with non-alcoholic fatty liver disease. *Panminerva medica* **56**, 189–193 (2014).
45. Rani, J. et al. T2diacod: a gene atlas of type 2 diabetes mellitus associated complex disorders. *Scientific Reports* **7**, 1–21 (2017).
46. Pressler, M. P., Horvath, A. & Entcheva, E. Sex-dependent transcription of cardiac electrophysiology and links to acetylation modifiers based on the gtex database. *Frontiers in Cardiovascular Medicine* **9** (2022).
47. Ortuso, F., Mercatelli, D., Guzzi, P. H. & Giorgi, F. M. Structural genetics of circulating variants affecting the sars-cov-2 spike/human ace2 complex. *Journal of Biomolecular Structure and Dynamics* 1–11 (2021).
48. Zhang, K. et al. gcastle: A python toolbox for causal discovery (2021). [arXiv:2111.15155](https://arxiv.org/abs/2111.15155).
49. Kalisch, M. & Buehlmann, P. Estimating high-dimensional directed acyclic graphs with the pc-algorithm (2005). [arXiv:math/0510436](https://arxiv.org/abs/math/0510436).
50. Consortium, G. O. The gene ontology resource: 20 years and still going strong. *Nucleic acids research* **47**, D330–D338 (2019).

## Acknowledgements

This work was funded by the Next Generation EU - Italian NRRP, Mission 4, Component 2, Investment 1.5, call for the creation and strengthening of 'Innovation Ecosystems', building 'Territorial R&D Leaders' (Directorial Decree n. 2021/3277) - project Tech4You - Technologies for climate change adaptation and quality of life improvement, n. ECS0000009. This work reflects only the authors' views and opinions, neither the Ministry for University and Research nor the European Commission can be considered responsible for them.

PV was partially supported by project SERICS (PE00000014) under the MUR385 National Recovery and Resilience Plan funded by the European Union: NextGen-erationEU.

## Author contributions

P.H.G. and F.G. conceived the rationale of the paper. P.V. led the bioinformatics design of the algorithms. A.D. designed the experimental setup, performed the experimental setup. All the authors analysed the results. A.D. wrote the draft version of the paper. P.V., A.D. and P.H.G. revised the paper and wrote the final version. All authors approved the manuscript.

## Declarations

### Competing interests

Authors declare that they have no competing interests.

### Additional information

Correspondence and requests for materials should be addressed to P.H.G.

Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024