Title

The effect of statistical normalization on network propagation scores.

Permalink
https://escholarship.org/uc/item/4k96j0mx

Journal
Bioinformatics, 37(6)

ISSN
1367-4803

Authors
Picart-Armada, Sergio
Thompson, Wesley K
Buil, Alfonso
et al.

Publication Date
2021-05-05

DOI
10.1093/bioinformatics/btaa896

Peer reviewed

Bioinformatics, 37(6), 2021, 845–852
doi: 10.1093/bioinformatics/btaa896
Advance Access Publication Date: 18 October 2020
Original Paper

Data and text mining

The effect of statistical normalization on network propagation scores

Sergio Picart-Armada [1,2,*], Wesley K. Thompson[3,4], Alfonso Buil[3] and Alexandre Perera-Lluna[1,2]

[1]B2SLab, Departament d'Enginyeria de Sistemes, Automàtica i Informàtica Industrial, Universitat Politècnica de Catalunya, CIBER-BBN, Barcelona, 08028, Spain, [2]Esplugues de Llobregat, Institut de Recerca Pediàtrica Hospital Sant Joan de Déu, Barcelona, 08950, Spain, [3]Mental Health Center Sct. Hans, 4000 Roskilde, Denmark and [4]Department of Family Medicine and Public Health, University of California, San Diego, La Jolla, CA, USA

*To whom correspondence should be addressed

Associate Editor: Jonathan Wren
Received on January 21, 2020; revised on September 18, 2020; editorial decision on October 1, 2020; accepted on October 7, 2020

Abstract

Motivation: Network diffusion and label propagation are fundamental tools in computational biology, with applications like gene–disease association, protein function prediction and module discovery. More recently, several publications have introduced a permutation analysis after the propagation process, due to concerns that network topology can bias diffusion scores. This opens the question of the statistical properties and the presence of bias of such diffusion processes in each of its applications. In this work, we characterized some common null models behind the permutation analysis and the statistical properties of the diffusion scores. We benchmarked seven diffusion scores on three case studies: synthetic signals on a yeast interactome, simulated differential gene expression on a protein–protein interaction network and prospective gene set prediction on another interaction network. For clarity, all the datasets were based on binary labels, but we also present theoretical results for quantitative labels.

Results: Diffusion scores starting from binary labels were affected by the label codification and exhibited a problem-dependent topological bias that could be removed by the statistical normalization. Parametric and non-parametric normalization addressed both points by being codification-independent and by equalizing the bias. We identified and quantified two sources of bias—mean value and variance—that yielded performance differences when normalizing the scores. We provided closed formulae for both and showed how the null covariance is related to the spectral properties of the graph. Despite none of the proposed scores systematically outperformed the others, normalization was preferred when the sought positive labels were not aligned with the bias. We conclude that the decision on bias removal should be problem and data-driven, i.e. based on a quantitative analysis of the bias and its relation to the positive entities.

Availability: The code is publicly available at https://github.com/b2slab/diffuBench and the data underlying this art icle are available at https://github.com/b2slab/retroData

Contact: sergi.picart@upc.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The guilt by association principle states that two proteins that interact with one another are prone to participate in the same, or related, cellular functions (Oliver, 2000). This cornerstone fact has motivated the exploration of network algorithms on interaction networks for protein function prediction (Sharan *et al.*, 2007). Network analysis has further proven its usefulness in other computational biology problems, such as prioritizing candidate disease genes (Barabási *et al.*, 2011), finding modular structures (Mitra *et al.*, 2013) and modelling organisms (Aderem, 2005).

Network propagation is a fundamental formalism to leverage network data in computational biology. Its theoretical basis revolves around graph spectral theory, graph kernels and random walks (Smola and Kondor, 2003). The central concept is that nodes carry abstract labels that, following the guilt by association principle, are propagated to the neighbouring nodes (Zoidi *et al.*, 2015). Unlabelled nodes can therefore be inferred a label based on the available data of their neighbours. Label propagation can be defined in several ways, such as the heat diffusion, the electrical model or random walks with restarts (RWR), some of which lead to equivalent formulations (Cowen *et al.*, 2017). While this article tackles classical propagation methods, there are more recent and sophisticated algorithms that can alleviate their shortcomings. Those include adaptive diffusion (Jiang *et al.*, 2017), non-linear diffusions for semi-supervised graph learning (Ibrahim and Gleich, 2019), graph convolutional neural networks (Sun *et al.*, 2020) and graph embeddings (Grover and Leskovec, 2016).

One of the most common diffusion formulations relies on the regularized Laplacian graph kernel (Smola and Kondor, 2003)—examples are provided throughout this paragraph. HotNet (Vandin et al., 2010) is a tool for finding modules with a statistically high number of mutated genes in cancer, after propagating the labels of mutated genes. The authors in Bersanelli *et al.* (2016) have found relevant modules from gene expression and mutation data, based on a diffusion process followed by an automatic subgraph mining. GeneMANIA (Mostafavi *et al.*, 2008) is a web server that predicts gene function by optimizing a combination of knowledge networks and running a diffusion process on the resulting network. TieDIE (Paull *et al.*, 2013) defines two diffusion processes in order to connect two sets of genes, applied to link perturbation in the genome with changes in the transcriptome. More generally, the predictive power of label propagation using graph kernels has been benchmarked in gene–disease association (Guala and Sonnhammer, 2017; Lee *et al.*, 2011; Valentini *et al.*, 2014).

Some studies have pointed out biases in diffusion scores and explored the effect of their removal. The authors of DADA (Erten *et al.*, 2011) have found that prioritization using RWR favours highly connected genes and suggest several normalization strategies. One of them computes a z-score that adjusts for the mean value and standard deviation estimated from propagation scores from random degree-preserving inputs. Another possibility is to normalize diffusion scores into empirical *P*-values, as used in the diffusion of *t*-statistics derived from gene expression (Cun and Fröhlich, 2013). The aim was to quantify robust biomarkers, whose diffusion score is unlikely to arise from a permuted input. In the discovery of enriched modules (Bersanelli *et al.*, 2016), the effect of the topology has been mitigated by combining diffusion scores with their empirical *P*-values. Similarly, exact z-scores and empirical *P*-values have been used for pathway analysis of metabolomics data (Picart-Armada *et al.*, 2017b). A recent study (Biran *et al.*, 2019) has normalized RWR into an empirical *P*-value, obtained from edge rewiring. Specifically, random degree-preserving networks have been built to re-run the propagation and draw values from the null distributions of scores. Another recent manuscript (Hill *et al.*, 2019) highlights biases in certain network propagation algorithms, related to the node degree.

Overall, a variety of measures to address the bias have emerged, but a systematic quantification and evaluation of the biases are missing. The normalization can potentially backfire, for instance by missing highly connected nodes that are associated with the property under study (Erten *et al.*, 2011). The goal of this manuscript is to provide a quantitative way to assess the presence of the bias and its alignment with the node labels, in order to understand the impact and adequateness of the normalization.

2 Approach

Here, we address the basic statistical properties of the normalization of single-network diffusion scores to remove topology-related biases. We define and quantify two sources of bias. Both are derived from a statistical standpoint, based on the exact means and variances of the null distributions of the diffusion scores under input permutation. Differences in mean values between nodes should be the first indicator of systematic advantages: nodes with the highest means will often be prioritized over those with the lowest means. In their absence, differences in variances should be examined instead, as nodes with highest spread can be more likely to reach extreme scores. We compare classical and normalized propagation, as implemented in diffuStats (Picart-Armada *et al.*, 2017a), in data with and without bias. The main results are derived for the commonly used regularized Laplacian kernel, although most of them apply to other graph kernels and, to a lesser extent, to random walks with restarts. Special emphasis is placed on identifying scenarios under which normalization is beneficial or detrimental and on understanding the underlying reasons why.

3 Methods

3.1 Diffusion scores

We include seven diffusion scores that are part of the diffuStats package (Picart-Armada *et al.*, 2017a): f_{raw}, f_{ml}, f_{gm}, f_{ber_s}, f_{mc}, f_z and f_{ber_p}. These scores are variations of the original diffusion model with a regularized unnormalized Laplacian kernel (Smola and Kondor, 2003). Labelled nodes are referred to as positives if they have the property of interest, and negatives otherwise.

3.1.1 Unnormalized scores

The starting point is the f_{raw} score, which requires a graph kernel K (Smola and Kondor, 2003) and input vector y_{raw} and is computed as:

$$f_{raw} = Ky_{raw} \tag{1}$$

This work focuses on the unnormalized, regularized Laplacian kernel for K, for being a widespread choice in the computational biology literature (electrical model, heat or fluid propagation). The values in y_{raw} reflect the weights of each type of node: 1 for positives and 0 for negative and unlabelled entities.f_{ml} and f_{gm} differ from f_{raw} by setting a weight of -1 on negative nodes. f_{gm} also weighs unlabelled nodes with a bias term adapted from GeneMANIA (not to be confused with the diffusion bias). On the other hand, f_{ber_s} measures the relative change between f_{raw} and y_{raw}, with a moderating parameter ϵ:

$$f_{ber_s}(i) = \frac{f_{raw}(i)}{y_{raw}(i) + \epsilon}. \tag{2}$$

3.1.2 Normalized scores

Normalized scores attempt to equalize nodes that systematically show low or high scores, regardless of the input and due to the specific topology of the network. The lynchpin of normalization is the null distribution of the diffusion scores under a random permutation π of the labelled nodes. The null scores arise from applying f_{raw} to a randomized input $X_y = \pi(y_{raw})$ and comparing, for the ith node, $f_{raw}(i)$ to its null distribution $X_f(i)$, where $X_f = KX_y$. An empirical P-value can be computed through Monte Carlo trials for the ith node on N trials:

$$p(i) = \frac{r_i + 1}{N + 1}, \tag{3}$$

where r_i is the number of randomized trials having an equal or higher diffusion score in node i. In order to assign high scores to relevant nodes, the score is defined as $f_{mc}(i) = 1 - p(i)$. We also include a parametric alternative to f_{mc} by computing z-scores for each node i:

$$f_z(i) = \frac{f_{raw}(i) - \mathrm{E}(X_f(i))}{\sqrt{\mathrm{Var}(X_f(i))}} \tag{4}$$

The expected value and variance of the null distributions are analytically determined (see Supplementary Material S1). Thus, f_z has a computational advantage over Monte Carlo trials.

Finally, a hybrid combining an unnormalized and a normalized score is provided, inspired by how (Bersanelli *et al.*, 2016) moderated the effect of hubs: f_{raw}: $f_{ber_p}(i) = -\log_{10}(p(i))f_{raw}(i)$.

3.2 Metrics and baselines

Two baseline methods were used. First pagerank (Page *et al.*, 1999), regarded as an input-naïve centrality measure (default damping factor of 0.85), to measure the predictive power of a basic network property. Second, a random predictor, to set an absolute baseline. Performances were quantified with two metrics: the area under the receiver operating characteristic curve (AUROC) and the area under the precision-Recall curve (AUPRC), as implemented in the precrec package (Saito and Rehmsmeier, 2017). For clarity, the ranking (ordering) of the nodes for any given score and instance was normalized to lie in $[0, 1]$ by dividing it by the number of ranked nodes, so that top suggestions corresponded to ranks close to 0.

3.3 Bias quantification

The reference expected value of the ith node $b^{\mathcal{K}}_{\mu}(i)$ (equation 7) was defined as proportional to the expected value of its null distribution $X_f(i)$ (equation 5). Reference expected values that vary across nodes can indicate systematic differences in the diffusion scores of such nodes.

In the absence of differences in the reference expected value, variance-related bias was analysed instead. The reference variance of the ith gene $b^{\mathcal{K}}_{\sigma^2}(i)$ (equation 8) was defined as, up to an additive constant, the base 10 logarithm of the variance of $X_f(i)$, straightforward to obtain from the covariance matrix (equation 6). The rationale is that the scores of nodes with varying dispersion measures should not be compared directly.

3.4 Performance explanatory models

Explanatory models have found use in the formal description of differences in performance as a function of design factors (Lopez-del Rio *et al.*, 2019; Picart-Armada *et al.*, 2019). Following (Picart-Armada *et al.*, 2019), the trends in AUROC and AUPRC were described through logistic-like quasibinomial models with a logit link function, as a generalization of logistic models to prevent over and under-dispersion issues.

Table 1 presents the main model for each case study. The categorical regressors were: method, metric (AUROC or AUPRC), biased (refers to the signal, true or false), strat (labelled, unlabelled or overall), array (ALL or Lym) and the parameters k, r and p_max for the second case study. path_var_ref was quantitative, equal to the reference pathway variance $bp^{\mathcal{K}}_{\sigma^2}$ (equation 9). The responses were either AUROC, AUPRC or both mixed, the latter denoted by performance.

4 Materials

The evaluation of the diffusion scores was performed on three datasets of different nature, as described in Table 1: (i) synthetic signals on a yeast interactome, (ii) pathway-based synthetic signals on a human network and (iii) real signals on another human network.

4.1 Networks

4.1.1 Yeast network

A small yeast network was used to demonstrate the casuistic of diffusion scores properties. Medium and high confidence interactions from several sources were provided by the original study (Von Mering *et al.*, 2002), as found in the igraphdata R package (Csardi, 2015). It contains 2617 proteins and 11 855 unweighted edges, but we worked only with its largest connected component (2375 proteins, 11 693 edges).

4.1.2 HPRD network

The diffuse large B-cell lymphoma study, available in the R package DLBCL (Dittrich and Beisser, 2010), contains a differential expression dataset accompanied by a human interactome network extracted from the Human Protein Reference Database (HPRD) (Mishra *et al.*, 2006). The original network encompasses 9385 proteins with 36 504 interactions, whose largest connected component (8989 nodes, 34 325 interactions) was extracted to compute the diffusion scores.

We derived two gene backgrounds based on expression arrays. The first background (Lym) was taken from the expression data from 2557 genes (2482 in the network) in the lymphoma study (Rosenwald *et al.*, 2002). The second background (ALL) was based on the acute lymphocytic leukaemia array (Chiaretti *et al.*, 2004), available in the ALL R package (Li, 2009), encompassing 6133 genes (5921 in the network).

4.1.3 BioGRID network

The Biological General Repository for Interaction Datasets (BioGRID) (Chatr-aryamontri *et al.*, 2017) is a public database with curated genetic and protein interaction from *Homo sapiens* and other organisms. BioGRID was retrieved in January 2017, but only keeping interactions dating from 2010 or older. The interactions were weighted according to (Cao *et al.*, 2014), under the assumptions that more publications about an interaction boost its confidence and that low-throughput technologies are more reliable that high-throughput ones. The network encompassed 11 394 nodes and 67 573 edges and was connected.

4.2 Datasets

4.2.1 Synthetic bias-based dataset

One-hundred biased and 100 unbiased instances of positive, negative and unlabelled nodes were generated in dataset (1) from Table 1, by sampling positive nodes with probabilities proportional to biased and unbiased scores. By construction, the frequencies of the positives drawn for biased signals were positively correlated with the reference expected value, whereas those of the unbiased signals were uncorrelated with it.

Nodes were partitioned into three equally sized pools, from which positive nodes were drawn: (i) labelled nodes that were fed to the diffusion methods, (ii) target nodes, the ones to be ranked and whose ground truth was known and (iii) filler nodes that were neither target nor labelled.

For each instance, a fixed fraction of labelled nodes x_e were uniformly sampled as positives, the rest of labelled nodes were deemed negatives and the target and filler nodes were left unlabelled. This input served two purposes: generate the ground truth in target nodes, and be the input for all the diffusion scores.

To generate the ground truth in target nodes of biased signals, the raw diffusion scores were computed from the input above. A fixed fraction of target nodes x_s was sampled with probabilities proportional to their raw scores, i.e. $p(i) \propto f_{raw}(i)$, to become positives.

Table 1. Case studies for characterizing biases and benchmarking diffusion scores

Case	Network	Positive nodes	Signal	Bias type	Purpose	Explanatory model for hypothesis testing
(1)	Yeast	Synthetic	Synthetic, bias-based	Mean value	Proof of concept	Performance $\sim$ method + method: biased + metric
(2)	HPRD	KEGG pathways	Pathway sub-sampling	Mean value	Background influence in bias	AUPRC $\sim$ method + method : strat + array + k + r + $p_{\max}$ + fdr
(3)	BioGRID	KEGG pathways	Prospective pathway prediction	Variance	Bias in a common scenario	AUROC $\sim$ method + method : $\text{path}_{\text{var}}\text{ref}$

Interactions in explanatory models are denoted by a colon.

The remaining target nodes would remain negatives, completing the ground truth. The regularized unnormalized Laplacian kernel is endorsed by physical models that ensure $f_{raw}(i) > 0$ provided that inputs have one or more positives and the graph is connected. Analogously, unbiased signals were generated by sampling a fraction of target nodes x_s, but with probabilities roughly proportional to the unbiased diffusion scores mc: $p(i) \propto f_{mc}(i) + \frac{1}{N+1}$. By definition, the frequency of appearance of the target nodes was independent of the bias, and the small offset ensured $p(i) > 0$.

In both cases, after sampling the ground truth, the same input was used again for all the diffusion scores, in order to rank the target nodes and compute the corresponding AUROC and AUPRC.

4.2.2 Pathway sub-sampling dataset

Synthetic gene expression statistics were generated, based on pathways in the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa *et al.*, 2017), and on two array-based gene backgrounds described within the HPRD network. Genes outside the background were hidden (unlabelled), and genes inside were given P-values for differential expression.

Each signal derived from k random KEGG pathways. The pathways were assumed to be affected as a whole, but only a sampled portion of r genes showed differential expression patterns. The P-values of the differential expressed genes were uniformly sampled from $[0, p_{max}]$, whereas the rest of genes were uniform in $[0, 1]$, following a previous study (Rajagopalan and Agarwal, 2005).

For both expression arrays, genes with an FDR $< 5\%$ or 10% within their background were used as positives, the remaining background genes as negatives and the hidden nodes were deemed unlabelled. Notice that, by definition, this procedure generated false positives and false negatives among the input genes.

The target genes were those belonging to the k affected pathways, including those with no apparent differential expression and those among the unlabelled nodes. Methods were compared using the AUROC and AUPRC, computed separately on labelled, unlabelled genes and overall, on a grid of parameters: $k \in \{1, 3, 5\}$, $r \in \{0.3, 0.5, 0.7\}$ and $p_{max} \in \{10^{-2}, 10^{-3}, 10^{-4}\}$. For each combination of parameters, $N = 50$ instances were simulated.

4.2.3 Prospective pathway dataset

The input lists consisted of the genes in 139 KEGG pathways from March 14, 2011. The target genes were the newly added genes in the same KEGG pathways in August 18, 2018 release. The 139 pathways had new genes in the latter release after mapping to the network.

AUROC and AUPRC were computed on each pathway, always excluding the input positive genes. The bias was examined at the pathway level, assessing whether the properties of their new genes differed from those of the rest of network genes. It was defined as the median reference variance of its new genes minus the median reference variance of all the genes besides old and new pathway genes (equation 9).

5 Results

5.1 Properties of diffusion scores

Some of the diffusion scores are equivalent in certain scenarios. In the absence of unlabelled nodes and using kernels based on the unnormalized graph Laplacian, f_{raw}, f_{ml} and f_{gm} lead to an identical node prioritization. More generally, the results using only two classes (and therefore two real values $y^+ > y^-$ as weights) always lead to the same ranking as f_{raw}. An analogous result holds for the weights of the positives and the unlabelled, $y^+ > y^u$, in the absence of negative nodes.

The normalized scores f_{mc} and f_z are invariant to changes in the weights of the positive and negative examples, regardless of the presence of unlabelled nodes and the graph kernel. This property simplifies the diffusion setup and leads to weight-independent results. Along with equations 5 and 6, this holds even if the matrix K in equation 1 is not a kernel, like the random walk similarity matrices in Cowen *et al.* (2017).

We also provide the closed form of the null expected value and covariance matrix of the raw scores, governed by the identifiers of the n_l labelled nodes (out of n). If $\mathcal{K}$ contains only their corresponding columns from K, and $\mathcal{Y}$ is the input vector y_{raw} restricted to them, then:

$$\mathbb{E}(X_f) = \mu_{\mathcal{Y}} \mathcal{K} 1_{n_l} \tag{5}$$

$$\Sigma(X_f) = \sigma^2_{\mathcal{Y}} \mathcal{K} M_{n_l} \mathcal{K}^T \tag{6}$$

$\mu_{\mathcal{Y}} = \frac{1}{n_l} \sum_{i=1}^{n_l} \mathcal{Y}_i$ and $\sigma^2_{\mathcal{Y}} = \frac{1}{n_l - 1} \sum_{i=1}^{n_l} (\mathcal{Y}_i - \mu_{\mathcal{Y}})^2$ are the mean and variance of the labels. $M_k = I_k - \frac{1}{k} 1_k 1_k^T$, being I_k the $k \times k$ identity matrix and 1_k the column vector with k ones.

If a graph kernel based on the unnormalized Laplacian is used, the covariance of the null distribution (equation 6) is closely related to the spectral properties of the labelled nodes. In particular, in the absence of unlabelled nodes, the leading eigenvector of the null covariance is, up to a sign change, the Fiedler-vector, commonly used for graph clustering (Smola and Kondor, 2003). The statistical normalization is therefore endowed with a topological basis. This sheds light on prior empirical observations that, even when the bias can relate to the node degree, there must be further topological factors involved (Hill *et al.*, 2019).

Because $\mu_{\mathcal{Y}}$ and $\sigma^2_{\mathcal{Y}}$ are multiplicative constants and inherent to the labels, the topology-related mean value and variance references of the ith node are defined as follows. We assume $n_l \geq 2$ because if $n_l \in \{0, 1\}$ there is nothing to permute.

$$b^{\mathcal{K}}_{\mu}(i) := [\mathcal{K} 1_{n_l}]_{i1} = \sum_{j=1}^{n_l} \mathcal{K}_{ij} \tag{7}$$

$$b^{\mathcal{K}}_{\sigma^2}(i) := \log_{10}\left(\left[\mathcal{K} M_{n_l} \mathcal{K}^T\right]_{ii}\right) = \log_{10}\left(\sum_{j=1}^{n_l}\left(\mathcal{K}_{ij} - \frac{b^{\mathcal{K}}_{\mu}(i)}{n_l}\right)^2\right). \tag{8}$$

Equation 5 implies that there are two scenarios free of the expected value bias: $\mu_{\mathcal{Y}} = 0$ (centred input), or $n_l = n$ and a kernel K based on the unnormalized Laplacian, rendering $b^{\mathcal{K}}_{\mu}$ constant (see Supplementary Material S1). The ith null variance (equation 6) can be exactly zero, either because $\sigma^2_{\mathcal{Y}} = 0$ (constant input), or because the topology forces $[\mathcal{K} M_{n_l} \mathcal{K}^T]_{ii} = 0$. In practice, the latter is expected to happen in small connected components without any labelled nodes. Both cases render the ith score constant, therefore lacking interest, and leave f_z undefined.

A dedicated analysis revealed that the statistical moments in the yeast and the BioGRID networks were affected by edge pre-filtering (Supplementary Material S5). When removing lower-confidence edges, $b^{\mathcal{K}}_{\mu}$ tended to increase in the labelled nodes and to decrease in the unlabelled ones ($p < 10^{-16}$ in six comparisons, two-sided paired Wilcoxon test), thus magnifying the differences between both.

In the retrospective dataset, the reference of a given pathway P, conceived to summarize its properties into a single number, was defined by subtracting the median reference of its new genes, new(P) to that of the genes that never belonged to it, others(P):

$$bp^{\mathcal{K}}_{\sigma^2}(P) := \underset{i \in \text{new}(P)}{\text{median}}\{b^{\mathcal{K}}_{\sigma^2}(i)\} - \underset{i \in \text{other}(P)}{\text{median}}\{b^{\mathcal{K}}_{\sigma^2}(i)\}. \tag{9}$$

The mathematical proofs of the properties and illustrative examples can be found in Supplementary Material S1.

5.2 Synthetic signals in yeast

5.2.1 Bias in diffusion scores

Supported by equation 5, the presence of unlabelled nodes originated different expected values among the nodes. We hypothesized that f_{raw} would be biased to favour nodes with high $b^{\mathcal{K}}_{\mu}$, whereas f_{mc} and f_z would prioritize in a more unbiased manner. Figure 1A

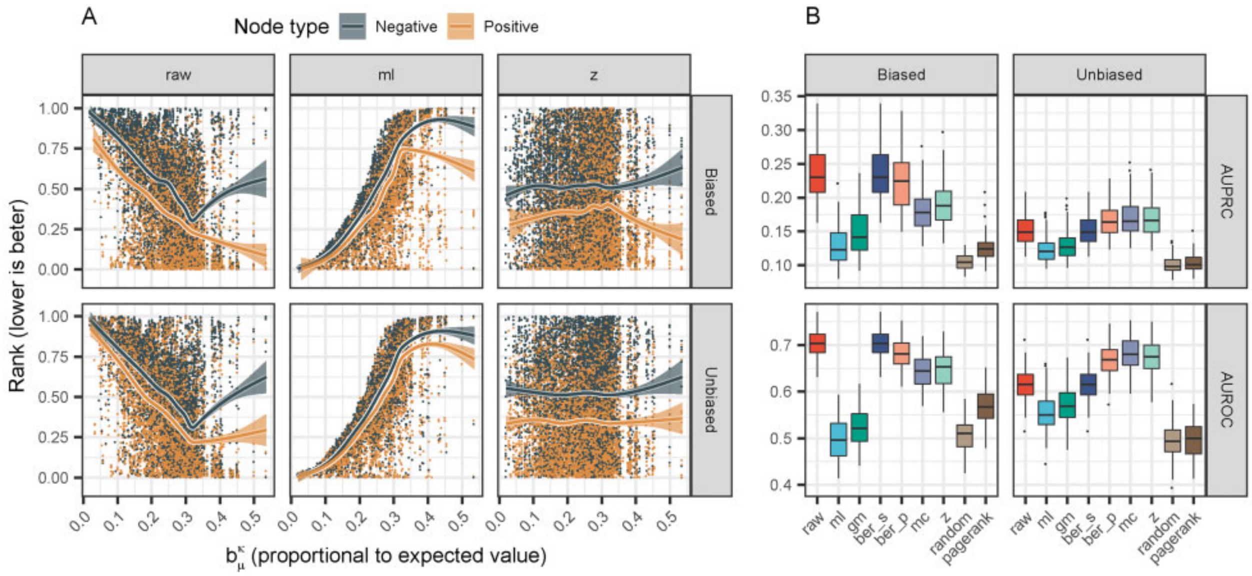

Fig. 1. Analysis of biased and unbiased synthetic signals on the yeast network. Nodes showed a mean value-related bias, see Supplementary Material S2. (**A**) Effects of the mean value bias in on the average node ranking, under biased and unbiased signals. Lines correspond to Generalized Additive Models with $y \sim s(x, bs = \grave{c}s'')$ and 0.95 confidence intervals. raw and ml tended to find positives with high and low $b_\mu^\mathcal{K}$, respectively. z found positives in a more uniform manner. (**B**) Performance in terms of AUROC and AUPRC. The lower and the higher hinges represent the first and third quartiles, with the median indicated by the intermediate bar. The whiskers extend up to 1.5 times the interquantile range from the box; more distant data points are displayed as outliers. raw was better suited for biased signals, for which the pagerank baseline also outperformed a random predictor. Conversely, z worked best on unbiased signals

confirms both trends. The data imbalance (negatives outnumbered positives in the input) had the opposite biasing effect on f_{ml}, favouring nodes with low $b_\mu^\mathcal{K}$.

5.2.2 Performance

In biased signals, target nodes with higher $b_\mu^\mathcal{K}$ were sampled as positives more often (see Supplementary Material S2), which (i) benefitted the unnormalized scores raw over z and (ii) endowed the pagerank baseline with predictive power. Unbiased signals led to a uniform density of positives across $b_\mu^\mathcal{K}$, which (iii) was better handled by z than by raw (Fig. 1B). Claims (i), (ii) and (iii) were statistically significant for AUROC and AUPRC (Tukey's method, FDR $< 10^{-10}$ in all cases, see Supplementary Material S2). Also, f_{ber_p} was a good compromise between raw and z.

Based on these results, we suggest a systematic criterion to choose whether to normalize in the general case, by assessing (i) the presence of the expected value-related bias by checking if $b_\mu^\mathcal{K}$ is constant among the nodes to be prioritized and (ii) the expected or hypothetical dependence between $b_\mu^\mathcal{K}$ and the labels to be predicted. In this proof of concept, differences in $b_\mu^\mathcal{K}$ bias were present and normalization was discouraged when $b_\mu^\mathcal{K}$ was aligned with the positives. If $b_\mu^\mathcal{K}$ is constant, $b_{\sigma^2}^\mathcal{K}$ should be examined instead, see the retrospective pathway dataset.

5.3 Simulated differential expression

5.3.1 Bias in diffusion scores

Analogously to the yeast dataset, the presence of unlabelled nodes led to differences in $b_\mu^\mathcal{K}$ among nodes, see Figure 2A. We hypothesized that the main source of bias would arise from such heterogeneity, i.e. that unnormalized scores would be prone to find positives among the highest expected values. In both arrays, the nodes belonging to one or more pathways had, compared with nodes outside, (i) larger $b_\mu^\mathcal{K}$ within the unlabelled genes, but (ii) lower $b_\mu^\mathcal{K}$ within the labelled nodes. Overall, (iii) labelled genes showed larger $b_\mu^\mathcal{K}$ than unlabelled genes. Figure 2A portrays the claims (i), (ii) and (iii) in both arrays—the six statements were significant with $p < 10^{-16}$, two-sided Wilcoxon test (see Supplementary Material S3).

5.3.2 Performance

The performance, as predicted by the explanatory models, was influenced by the background used to compute the metrics, especially for AUPRC. Taking as reference f_{raw} and f_z, raw performed best in the unlabelled background and overall whereas z was preferable in the labelled background (Fig. 2B). The three claims were significant in both arrays (Tukey's method, $p < 10^{-10}$, see Supplementary Material S3).

Differences in performance were consistent with the expected value-related bias: potential positives suffered from lower $b_\mu^\mathcal{K}$ in the labelled genes and benefitted from greater $b_\mu^\mathcal{K}$ in the unlabelled part. In views of this, the natural choices were z and raw, respectively.

To understand why raw outperformed z in overall performance, note how by hypothesis the top candidates from raw should come from the labelled genes due to their high $b_\mu^\mathcal{K}$ against the unlabelled genes, whilst z should equalize predictions from both backgrounds. Predictions from the labelled part were more reliable owing to the presence of prior data on the genes (Fig. 2B). z equalized both backgrounds, shuffling reliable and unreliable predictions, and undermined overall performance.

Finally, an indirect assessment of the bias (PageRank centrality) fell short to explain performance differences in (i) and suggested that biased scores were preferrable in the three cases, see Supplementary Material S3. This highlights the importance of using a precise quantification of the bias.

5.4 Prospective pathway prediction

5.4.1 Bias in diffusion scores

Here, $b_\mu^\mathcal{K}$ was constant among all the nodes, as a consequence of using the unnormalized Laplacian without unlabelled nodes (see Supplementary Material S1). Differences still existed in terms of $b_{\sigma^2}^\mathcal{K}$ (Fig. 3A), implying that the normalization would make a difference.

However, the interpretation of the normalization impact was not as straightforward as for the expected value bias. With the paradigm of the z-scores z, deviations from the expected value exacerbate under small variances and shrink under large variances. Notice how this does not imply the natural hypothesis that nodes with larger variances (respectively smaller) must drop (respectively rise) in the ranking, because ranking modifications take place around the mean.

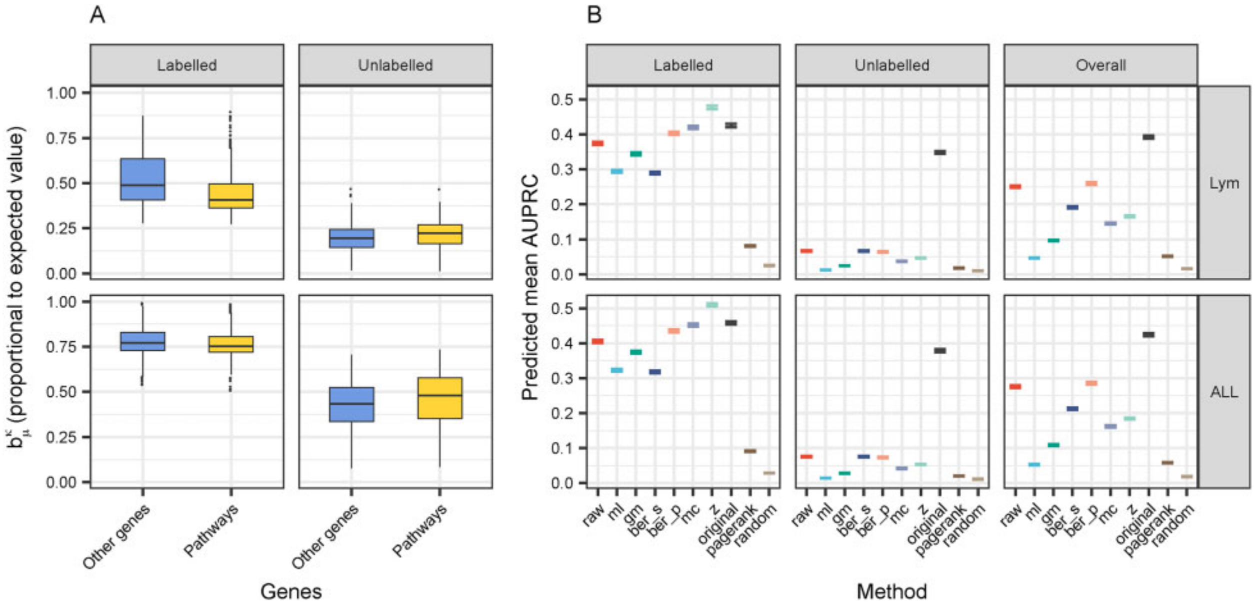

Fig. 2. Performance in the DLBCL dataset. (**A**) Expected value-related bias. Within the labelled genes of both arrays, those in pathways had lower $b_\mu^{\mathcal{K}}$ that those outside. Within the unlabelled genes, this tendency was inverted. Overall, labelled genes had higher $b_\mu^{\mathcal{K}}$ than unlabelled genes. (**B**) Predicted AUPRC (0.95 confidence interval) using the explanatory model in Table 1 and Supplementary Material S3. Besides diffusion scores, three baselines were included: original (ranking by the *P*-values), pagerank and random. In both arrays (*ALL* and *Lym*), raw outperformed z in unlabelled nodes and overall, while z was preferable in the labelled genes

Figure 3B reflects how z actually recovered more high-variance positive nodes than raw.

Similarly to prior observations from Figure 1A, the normalized scores tended to find the positives in a less-biased manner. Positive nodes with a high variance were rarely found by raw, whereas z distributed them more evenly along the ranking (Figure 3B). This improvement came at the cost of missing positives with lower variances.

5.4.2 Performance

The properties of the diffusion scores helped simplify this case study, as f_{ml}, f_{gm} and f_{ber_s} were left out for being redundant with f_{raw}. f_{ml} and f_{gm} for using the unnormalized Laplacian without unlabelled nodes, and f_{ber_s} because the genes to be prioritized were always labelled as negative in the input (see Corollary 1 and Proposition 3 in Supplementary Material S1).

The prospective prediction of pathway genes was a challenging task, given the low predicted AUPRCs for all the methods (see Supplementary Material S4). On the other hand, AUROC conveyed a richer view of the differences between methods. The explanatory model (Fig. 3C and D) showed that unnormalized scores were more affected by the presence of bias, reflected in the larger magnitude of their interaction terms (−1.387 for raw against −0.484 for z, $p < 10^{-4}$, Tukey's method). Overall, the casuistic among the bias of new pathway genes favoured z over raw (FDR $= 5.39 \cdot 10^{-9}$, two-sided paired Wilcoxon test). This conclusion did not apply to early retrieval, as it could not be proven for AUPRC (FDR $= 0.701$).

The negative sign of the interaction terms was also insightful: all the proper methods encountered more difficulties in finding loosely connected genes. This was expected, since there is less network data involving such genes, translating into unreliable predictions.

The impact of removing lower confidence edges before the propagation was explored in Supplementary Material S4. Moderate and aggressive filtering strategies (confidence thresholds of 0.3 and 0.9) created isolated nodes and lowered the AUROC of raw and z, justifying the default option of no filtering. Without accounting for the sign, the impact of deciding to normalize was comparable to that of switching to the aggressive filtering (95% confidence intervals of [0.502, 0.747] and [−0.73, −0.432] in logit scale). This suggests that considering the statistical normalization should be on par with other standard decisions.

6. Conclusion

In this study, we ratified that diffusion scores are biased due to the graph topology. We introduced two direct quantifications of the bias, in terms of the expected value and variance of the null distribution of the diffusion scores under input permutation. We analysed the benefits and pitfalls of using unbiased, statistically normalized scores and discussed several choices of the label weights when defining the diffusion process.

We proved equivalences between scores under certain conditions, helping simplify the setup of the diffusion, and discovered that normalized alternatives are invariant under label weights changes. We found an explicit link between principal directions of the null covariance and the spectral features of the network.

We applied the diffusion-based prioritization on three scenarios: two with a mean value-related bias and one with a variance-related bias. Class imbalance and node topology had an impact in unnormalized scores, whereas normalized scores were more robust to both phenomena given their weight-independent definition. The parametric normalization requires no permutations compared to Monte Carlo trials and performed equally or better, providing a convenient way to normalize. While mean value bias was straightforward to characterize, variance bias was less intuitive albeit of noticeable impact. In general terms, the statistical normalization is advised if the positives are not aligned with the bias, and discouraged otherwise. The statistical background, i.e. which nodes are permuted, is a key piece that should be clearly stated in every application. Bias assessment should be carried through its direct quantification instead of indirect indicators, which can be misleading.

We conclude that the statistical normalization can have a noticeable impact, be beneficial or detrimental, and the decision should follow from the dependence between the node bias and the hypothetical or desired properties of the new positives. Topology-related

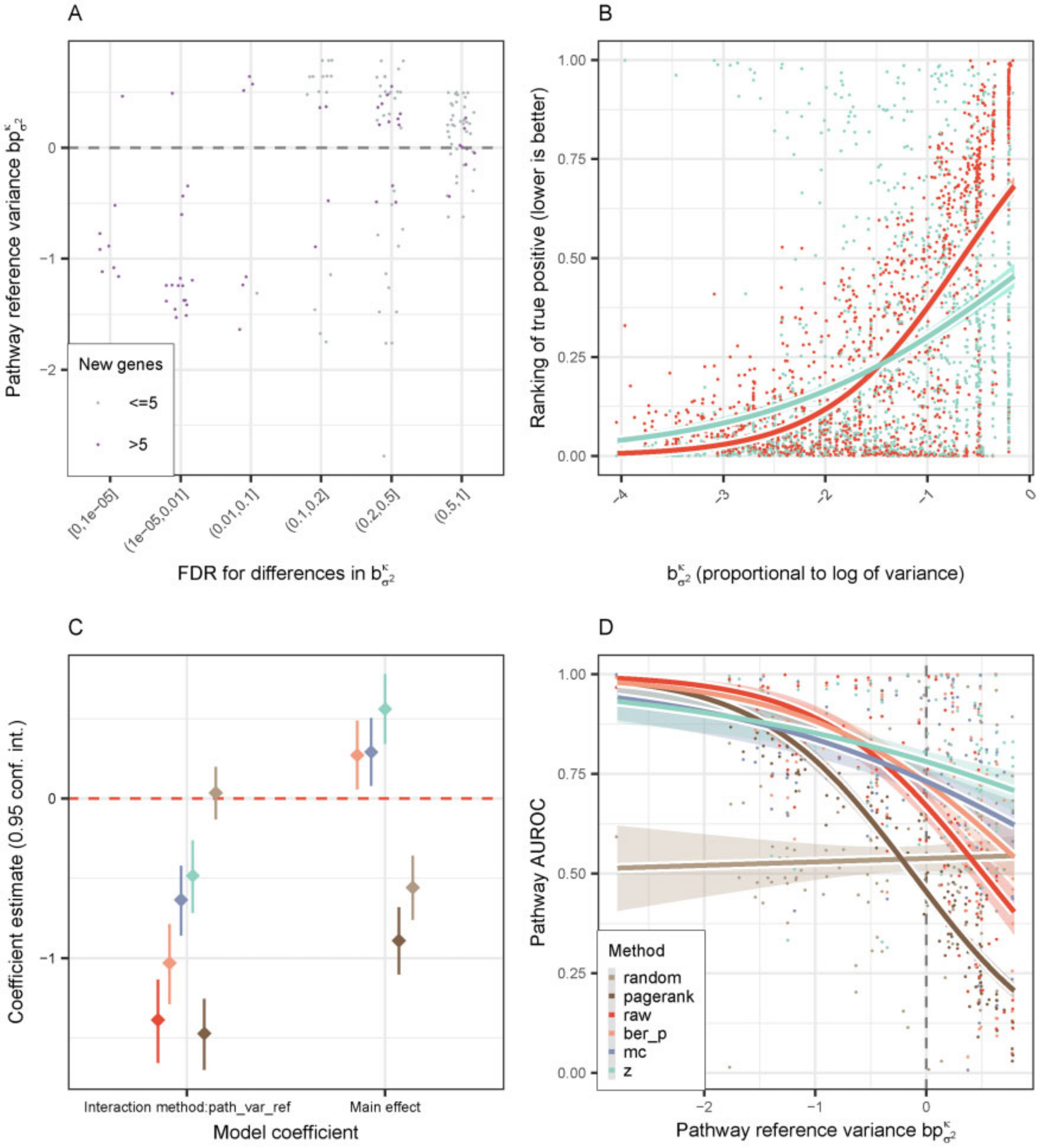

Fig. 3. Analysis of the prospective dataset. (**A**) Pathway-wise comparison of new genes against the remaining genes outside the pathway, in terms of $b^{\kappa}_{\sigma^2}$. Several pathways showed significant differences in both directions (two-sided Wilcoxon test). The *x* axis was jittered for clarity. (**B**) Ranking of the positives using raw and z. Each data point is the relative ranking of a positive gene in one of the pathways, i.e. before computing pathway-level metrics. Lines correspond to a quasi-logistic fit with a 0.95 confidence interval. raw scores were more sensitive at low standard deviations, whereas z stood more uniform. (**C**) Coefficients of the model AUROC $\sim$ method + method : path$_{\text{var}}$ref with a 0.95 confidence interval, where the interaction term involved the variance bias. The main effect of raw was not depicted because it was the reference level of method. (**D**) Predicted AUROC across all the pathways, as a function of the bias. z was less sensitive to the bias, due to its interaction term in (C) being closer to 0. Lines correspond to a quasi-logistic fit with a 0.95 confidence interval

bias can manifest in different ways (mean value- or variance-related bias) and each instance should be properly characterized.

Acknowledgements

S.P. thanks Guillem Belda-Ferrín for reviewing the mathematical proofs. S.P. thanks Imanol Morata Martínez and Camellia Sarkar for fruitful discussions and useful suggestions. A.P. and S.P. thank CIBERDEM and CIBER-BBN for funding, both initiatives of the Spanish ISCIII. S.P. thanks the AGAUR FI-scholarship programme.

Funding

This work was supported by the Spanish Ministry of Economy and Competitiveness (MINECO) [TEC2014-60337-R and DPI2017-89827-R to A.P.] and the National Institutes of Health (NIH) [R01GM104400 to W.T.]. A.P and S.P thank the Networking Biomedical Research Centre in the subject area of Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN), initiatives of Instituto de Investigación Carlos III (ISCIII), and Share4Rare project (Grant Agreement 780262). B2SLab is certified as 2017 SGR 952.

Conflict of Interest: none declared.

References

Page,L. et al. (1999). *The PageRank Citation Ranking: Bringing Order to the Web. Stanford InfoLab, Stanford*, CA, USA.

Aderem,A. (2005) Systems biology: its practice and challenges. *Cell*, **121**, 511–513.

Barabási,A.-L. *et al.* (2011) Network medicine: a network-based approach to human disease. *Nat. Rev. Genet*., **12**, 56–68.

Bersanelli,M. *et al.* (2016) Network diffusion-based analysis of high-throughput data for the detection of differentially enriched modules. *Sci. Rep*., **6**, 34841.

Biran,H. *et al.* (2019) Comparative analysis of normalization methods for network propagation. *Front. Genet*., **10**, 4.

Cao,M. *et al.* (2014) New directions for diffusion-based network prediction of protein function: incorporating pathways with confidence. *Bioinformatics*, **30**, i219–i227.

Chatr-Aryamontri,A. *et al.* (2017) The biogrid interaction database: 2017 update. *Nucleic Acids Res*., **45**, D369–D379.

Chiaretti,S. *et al.* (2004) Gene expression profile of adult t-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. *Blood*, **103**, 2771–2778.

Cowen,L. *et al.* (2017) Network propagation: a universal amplifier of genetic associations. *Nat. Rev. Genet*., **18**, 551–562.

Csardi,G. (2015). *igraphdata: A Collection of Network Data Sets for the 'igraph' Package*. R package version 1.0.1. Available at: http://CRAN.R-project.org/package=igraphdata (October 2020, date last accessed).

Cun,Y., and Fröhlich,H. (2013) Network and data integration for biomarker signature discovery via network smoothed T-statistics. *PLoS One*, **8**, e73074.

Dittrich,M., and Beisser,D. (2010). *DLBCL: diffuse large B-cell lymphoma expression data*. R package version 1.16.0. Available at: doi: 10.18129/B9.bioc.DLBCL (October 2020, date last accessed).

Erten,S. *et al.* (2011) DADA: degree-aware algorithms for network-based disease gene prioritization. *BioData Mining*, **4**, 19.

Grover,A. and Leskovec,J. (2016). node2vec: scalable feature learning for networks. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 855–864.

Guala,D., and Sonnhammer,E.L. (2017) A large-scale benchmark of gene prioritization methods. *Sci. Rep*., **7, 46598**.

Hill,A. *et al.* (2019) Benchmarking network algorithms for contextualizing genes of interest. *PLoS Comput. Biol*., **15**, e1007403.

Ibrahim,R. and Gleich,D. (2019). Nonlinear diffusion for community detection and semi-supervised learning. In: *The World Wide Web Conference*, pp. 739–750.

Jiang,B. *et al.* (2017) Aptrank: an adaptive pagerank model for protein function prediction on bi-relational graphs. *Bioinformatics*, **33**, 1829–1836.

Kanehisa,M. *et al.* (2017) Kegg: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res*., **45**, D353–D361.

Lee,I. *et al.* (2011) Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res*., **21**, 1109–1121.

Li,X. (2009). *ALL: A data package*. R package version 1.20.0. Available at: doi:10.18129/B9.bioc.ALL (October 2020, date last accessed).

Lopez-del Rio,A. *et al.* (2019) Evaluation of cross-validation strategies in sequence-based binding prediction using deep learning. *J. Chem. Inform. Model*., **59**, 1645–1657.

Mishra,G.R. *et al.* (2006) Human protein reference database 2006 update. *Nucleic Acids Res*., **34**, D411–D414.

Mitra,K. *et al.* (2013) Integrative approaches for finding modular structure in biological networks. *Nat. Rev. Genet*., **14**, 719–732.

Mostafavi,S. *et al.* (2008) GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biol*., **9**, S4.

Oliver,S. (2000) Guilt-by-association goes global. *Nature*, **403**, 601–603.

Paull,E.O. *et al.* (2013) Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE). *Bioinformatics*, **29**, 2757–2764.

Picart-Armada,S. *et al.* (2017a) diffuStats: an R package to compute diffusion-based scores on biological networks. *Bioinformatics*, **34**, 533–534.

Picart-Armada,S. *et al.* (2017b) Null diffusion-based enrichment for metabolomics data. *PLoS One*, **12**, e0189012.

Picart-Armada,S. *et al.* (2019) Benchmarking network propagation methods for disease gene identification. *PLoS Comput. Biol*., **15**, e1007276.

Rajagopalan,D., and Agarwal,P. (2005) Inferring pathways from gene lists using a literature-derived network of biological relationships. *Bioinformatics*, **21**, 788–793.

Rosenwald,A. *et al.* (2002) The use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma. *N. Engl. J. Med*., **346**, 1937–1947.

Saito,T., and Rehmsmeier,M. (2017) Precrec: fast and accurate precision–recall and ROC curve calculations in R. *Bioinformatics*, **33**, 145–147.

Sharan,R. *et al.* (2007) Network-based prediction of protein function. *Mol. Syst. Biol*., **3**, 88.

Smola,A.J., and Kondor,R. (2003). Kernels and regularization on graphs. In: Schölkopf, B., Warmuth, M.K. (eds.) *Learning Theory and Kernel Machines*. Springer, Berlin, Heidelberg, pp. 144–158.

Sun,M. *et al.* (2020) Graph convolutional networks for computational drug development and discovery. *Brief. Bioinformatics*, **21**, 919–935.

Valentini,G. *et al.* (2014) An extensive analysis of disease-gene associations using network integration and fast kernel-based gene prioritization methods. *Artif. Intell. Med*., **61**, 63–78.

Vandin,F. *et al.* (2010) Algorithms for detecting significantly mutated pathways in cancer. *Lect. Notes Comput. Sci*., 506–521.

Von Mering,C. *et al.* (2002) Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, **417**, 399–403.

Zoidi,O. *et al.* (2015) Graph-based label propagation in digital media: a review. *ACM Comput. Surveys*, **47**, 1–35.