Bioinformatics, 38(17), 2022, 4145–4152
https://doi.org/10.1093/bioinformatics/btac446
Advance Access Publication Date: 6 July 2022
Original Paper

Systems biology

Biological Random Walks: multi-omics integration for disease gene prioritization

Michele Gentili[1,†], Leonardo Martini[1,†], Marialuisa Sponziello [2,*] and Luca Becchetti[1,*]

[1]Department of Computer, Control, and Management Engineering Antonio Ruberti, Sapienza University of Rome, Rome, Italy and [2]Translational and Precision Medicine Department, Sapienza University of Rome, Rome, Italy

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Can Alkan

Received on November 17, 2021; revised on June 22, 2022; editorial decision on June 28, 2022; accepted on July 5, 2022

Abstract

Motivation: Over the past decade, network-based approaches have proven useful in identifying disease modules within the human interactome, often providing insights into key mechanisms and guiding the quest for therapeutic targets. This is all the more important, since experimental investigation of potential gene candidates is an expensive task, thus not always a feasible option. On the other hand, many sources of biological information exist beyond the interactome and an important research direction is the design of effective techniques for their integration.
Results: In this work, we introduce the Biological Random Walks (BRW) approach for disease gene prioritization in the human interactome. The proposed framework leverages multiple biological sources within an integrated framework. We perform an extensive, comparative study of BRW's performance against well-established baselines.
Availability and implementation: All codes are publicly available and can be downloaded at https://github.com/LeoM93/BiologicalRandomWalks. We used publicly available datasets, details on their retrieval and preprocessing are provided in the Supplementary Material.
Contact: marialuisa.sponziello@uniroma1.it or becchetti@diag.uniroma1.it
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

In recent years, through the advent of big data, genomics and quantitative *in silico* methodologies, medicine is witnessing tremendous advancements toward the understanding of the human pathophysiology (Silverman *et al.*, 2020). Gene–disease associations have been identified by genome-wide association studies (Hardy and Singleton, 2009) and more recently by whole exome or whole-genome sequencing studies (Koboldt *et al.*, 2013). While many of the mechanisms underlying these associations remain largely unclear, a growing body of research highlights associations between groups of interacting proteins and diseases within the so-called 'human interactome', representing the cellular network of all physical molecular interactions (Barabási *et al.*, 2011). A key feature is that disease proteins do not appear to be uniformly scattered across the interactome (Menche *et al.*, 2015), but they are prone to participation in common biological activities such as, e.g. genome maintenance, cell differentiation or growth signaling, which are the most relevant pathways in carcinogenesis (Ozturk *et al.*, 2018). For these reasons, while traditional single protein (i.e. magic bullet) approaches have limited effectiveness in addressing complex diseases, network-based ones can prove useful in identifying disease modules within the interactome, hopefully providing insights into key mechanisms and guiding the quest for therapeutic targets. Moreover, experimental investigation of potential gene candidates is an expensive task, thus not always a feasible option.

The human interactome refers to all protein–protein interactions (PPIs) within a cell, including regulatory interaction of transcription factors, metabolic enzyme-coupled interactions, protein complexes and kinase/substrate interactions. This network is largely incomplete. Currently, more than 140 000 interactions involving over 13 000 proteins are known [e.g. see Korcsmaros *et al.* (2017) and Gustafsson *et al.* (2014)]. The interactome-based approach to network medicine (Barabási *et al.*, 2011) proved effective for a number of diseases, e.g. by identifying putative biomarkers and subtypes, thus allowing a principled approach to drug targeting (Barabási *et al.*, 2011; Ozturk *et al.*, 2018). The need for new disease genes (or disease proteins) as putative candidates for diagnosis, prognosis or treatment, motivated the development of a number of algorithms for disease genes and modules prediction (Ghiassian *et al.*, 2015).

Two related classes of methodologies have emerged over the last decade as the most promising: module-based (Barabási *et al.*, 2011; Ghiassian *et al.*, 2015) and network propagation (Cowen *et al.*, 2017; Köhler *et al.*, 2008) algorithms. Module-based algorithms find topological, functional or disease modules in the interactome network, on the hypothesis that these represent cellular components likely involved in the same disease. Network propagation (or diffusion-based) algorithms leverage the information flow through nearby proteins in the network from initial (known) disease genes as their main ingredient.

While important and effective in many cases, the interactome is only one of many sources of biological information. An important research direction is the quest for effective techniques that allow the seamless integration of rich and heterogeneous biological sources into methods that were originally designed to leverage the topological features of the interactome (De Bie *et al.*, 2007; Dimitrakopoulos *et al.*, 2018; Shang and Liu, 2020).

Our contribution. In this work, we introduce the Biological Random Walks (BRW) framework for disease gene prioritization. The proposed framework leverages the integration of multiple biological sources within a propagation-based approach. We compare BRW's performance against well-established baselines, such as: RWR (Navlakha and Kingsford, 2010), DIAMOnD (Ghiassian *et al.*, 2015), DADA (Erten *et al.*, 2011) and RWR-M (Valdeolivas *et al.*, 2019). In particular, we investigate BRW's performance along different axes: (i) an in-depth, comparative analysis on four cancer phenotypes (i.e. breast cancer, lung adenocarcinoma, papillary thyroid cancer and colorectal adenocarcinoma); (ii) a broad comparative analysis of BRW's prioritization performance over a wide spectrum of Mendelian diseases with different characteristics and prior information available; (iii) an external validation using Food and Drug Administration (FDA)-approved drugs for breast cancer treatment, along with an evaluation of the algorithm results' stability across multiple population studies [some of the ideas presented in this submission, albeit in preliminary form and accompanied by a minimal, internal validation, appeared in Gentili *et al.* (2019). For the sake of completeness, Supplementary Section C.4 presents a comparison of our approach with Gentili *et al.* (2019)].

2 Materials and methods

2.1 BRW

BRW build on the hypothesis that integrating different biological information sources may better reflect the complexity of protein interactions in a cell's process. In light of this insight, our approach integrates information on pairwise protein interaction of the PPI network (Barabási *et al.*, 2011) with other biological data in a unified framework. Our approach is to some extent agnostic to the particular biological data source, as long as it affords a principled notion of similarity between proteins. In the remainder, we use bold lowercase to denote vectors and capital, non-bold letters to denote matrices. Given a vector $\mathbf{x}$, x_i denotes its i-th entry.

Notation. We represent the PPI as an undirected graph $G = (V, E)$, with genes as vertices. Any edge (i, j) represents a known PPI recorded in the PPI. We assume $|V| = n$ in the following. For a given gene i, $N_h(i)$ denotes the subset of G's vertices whose shortest path distance from i is exactly h.

2.1.1 Random walks with restart

Random Walk with Restart (RWR) (Köhler *et al.*, 2008) is a diffusion-based method, whose purpose is identifying disease modules that are topologically 'close' to known disease genes in the interactome. It was shown to outperform other prioritization algorithms in many cases (Navlakha and Kingsford, 2010). In a nutshell, this algorithm can be seen as performing multiple random walks over the PPI network, each starting from a *seed node* associated to a known disease gene, iteratively moving from one node to a random neighbor, thus simulating the diffusion of the disease phenotype across the interactome. More formally, the RWR is defined as:

$$\mathbf{p}^{(t+1)} = (1 - r)W\mathbf{p}^{(t)} + r\mathbf{q}. \quad (1)$$

Here, W is the column-normalized adjacency matrix of the graph and $\mathbf{p}^{(t)}$ is a vector, whose i-th entry $p_i^{(t)}$ is the probability of the random walk being at node i at the end of the t-th step. $r \in (0, 1)$ is the restart probability, i.e. the probability that the random walk is restarted from one of the (disease-associated) seed nodes in the next step. Upon a restart, the probability of restarting the random walk from some seed node j is q_j. Vector $\mathbf{q}$ is normally called a *personalization vector* in the Data Science literature. This random walk corresponds to an ergodic Markov chain (Levin and Peres, 2017) that admits a stationary distribution (i.e. a fixed point) $\mathbf{p}$. Nodes of the PPI are simply ranked by considering the corresponding entries of $\mathbf{p}$ in descending order of magnitude.

2.1.2 Biological information-aware random walks

For the sake of exposition, in the remainder, we refer to the biological information associated to a gene (e.g. the set of its annotations) as the set of its *features*. These can include (more precisely, be derived from) annotations from the Gene Ontology database (The Gene Ontology Consortium, 2019) (GO in the remainder) or gene expression levels. We remark that, in principle, any reliable information source on gene biology can be integrated. BRW ranks genes according to the main steps outlined below.

Unlike Köhler *et al.* (2008) and similar approaches, our method consists of two main steps: (i) extracting statistically significant features from biological data, using them to compute a personalization vector and a transition matrix used by the BRW algorithm; and (ii) using the stationary distribution of the corresponding random walk to rank genes. Our approach to computing aggregated personalization vectors and transition matrices is outlined below, with further details given in Supplementary Sections B.3 and B.4.

Computing a personalization vector. Both gene annotation data and gene expression levels allow derivation of personalization vectors that reflect some notion of similarity between a gene and a disease, with the latter represented by the set of its seed genes. In the first case, this similarity is defined in terms of knowledge about genes' involvement in various biological functions and/or diseases. In the second case, similarity is defined in terms of co-expression levels of different genes in subject cases as opposed to expression levels in a control group, using data about a population of patients involved in a clinical trial.

For annotation data, assume we have ℓ sources of biological information (e.g. GO, KEGG pathways etc.). Let S denote the seed set. For every $j = 1, \ldots, \ell$, we use $\mathcal{F}^j$ to denote the subset of annotations from the j-th source that are associated with at least one gene in S. Then, for every $j = 1, \ldots, \ell$, we select a subset of annotations $\hat{\mathcal{F}^j}$, filtering out annotations that are not statistically significant, as shown in Supplementary Figure S2 (i.e. $P - value > 10^{-5}$, using Fisher Exact Test and FDR correction), so that $\mathcal{F} = \cup_{j=1}^{\ell} \hat{\mathcal{F}^j}$ denotes the set of all statistically significant annotations that are associated with genes in S. Likewise, for every gene i (not necessarily belonging to S), we denote by $\mathcal{A}(i)$ the set of its annotations, possibly extracted from multiple biological information sources. We assign each gene i a weight θ_i, which reflects the extent to which i shares annotations that are statistically significant for genes that belong to the seed set of the disease under consideration. While other choices are possible, the definition we adopted reflects the extent of the *inclusion* of $\mathcal{A}_i$ in each of the ℓ sources (Fig. 1):

$$\theta_i = \sum_{j=1}^{\ell} \frac{|\mathcal{A}(i) \cap \hat{\mathcal{F}^j}|}{|\hat{\mathcal{F}^j}|}, \quad (2)$$

[it should be noted that Köhler *et al.* (2008) correspond to choose $\theta_i = 1$ if $i \in S$, $\theta_i = 0$ otherwise].

At this point, the components of a personalization vector $\mathbf{q}$ can be computed as follows:

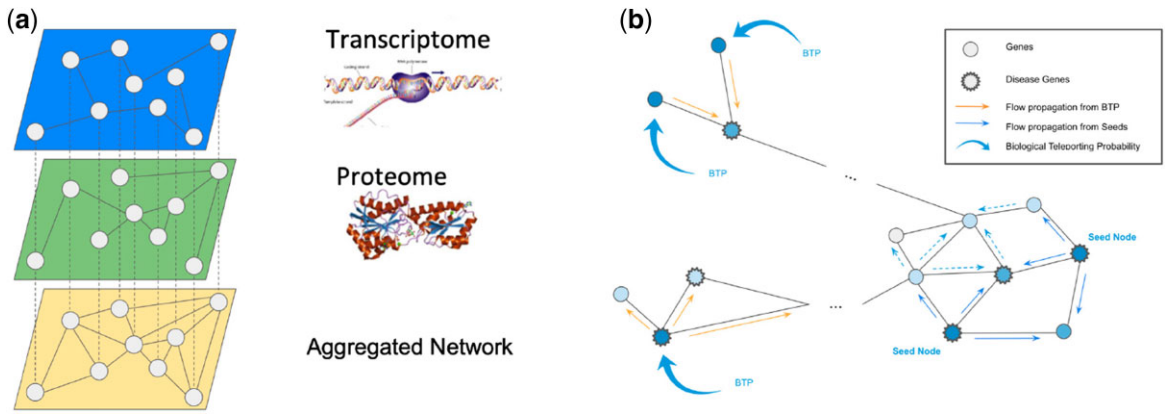

Fig. 1. Network aggregation and network flow of BRW. (**a**) Shows the preprocessing step that combines PPI and CO-Expression topology to derive a combined network. The transition matrix of PPI and CO-Expression network are combined using a convex combination to compute an aggregated transition matrix. Note that what is termed 'aggregated network' in tab (a) is actually a *weighted* network that corresponds to the aggregated transition matrix. (**b**) Shows the flow of the random walker when the personalization vector is biased with disease-specific information. In this way the flow can propagate from seed nodes (i.e. known disease genes), but also from node that are not part of the seed set but biologically similar to it (i.e. biological teleporting probability)

$$q_i = \frac{\theta_i}{\sum_{i=0}^{N} \theta_i}, \tag{3}$$

with N the number of genes we consider. Note that q_i denotes the probability that, upon teleportation, the random walkers jump to gene i. Further details are given in Supplementary Section B.3.1.

We next discuss how to compute personalization vectors from gene expression data. An important goal is the identification of differentially expressed (DE) genes, whose expression levels systematically differ between case (breast cancer cells) and a control group (normal breast cells). We follow the approach proposed in Menche *et al.* (2017), in which subjects of the control group are assigned personalized perturbation profiles (PEEPs), from which gene expression-aware personalization vectors can be derived. Succinctly put, for each gene i and for each subject j, the expression level l_i^j of gene i in subject j is compared with the distribution of the expression level of gene i within the control group by taking the corresponding z-score z_{ij}. This approach allows association of a 'bar code' to each subject. Following Menche *et al.* (2017), we set $|z_{ij}| > 2.5$ as the threshold to declare gene i DE in subject j. Following the general intuition stated in the introduction that disease genes generally are not scattered across the interactome, we also bias our choice toward DE genes that are closer to disease genes in the PPI. Eventually, we obtain a personalization vector $\mathbf{q}$ that reflects both genes' differential expressions and vicinity to disease genes. For full details on computing PEEPs and deriving gene expression-aware personalization vectors, we refer the reader to Supplementary Section B.3.1.

Computing a transition matrix. Similar approaches can be used to derive a transition matrix for the RWR. Both approaches rely on the PPI, differing on the way PPI's edges are assigned weights that reflect genes' similarity and determine the probabilities of edge traversals. We leverage categorical, biological information (e.g. gene annotations) by defining a weighted transition matrix W, in which each entry W_{ij} depends on the extent to which nodes/genes i and j of the PPI share common annotations (i.e. they are involved in common biological processes) that are also significant for the disease. In more detail, considered genes i and j, we define the following *Disease Specific Interaction Function* (DSI function in the remainder):

$$DSI(i,j) = |\mathcal{A}(i) \cap \mathcal{A}(j) \cap \mathcal{F}|, \tag{4}$$

where we remind that $\mathcal{A}(k)$ denotes the set of gene k's annotations, while $\mathcal{F}$ denotes the overall set of annotations that are statistically significant for disease genes. Intuitively, the higher $DSI(i, j)$, the more i and j share annotations that are also statistically significant for the disease under consideration. In the end, edges in the PPI will be assigned weights depending on the DSI function as follows:

$$W_{ij} = \begin{cases} c + DSI(i,j) & \text{if } (i,j) \in E, \\ 0 & otherwise. \end{cases} \tag{5}$$

Here, c is a positive constant that accounts for usual sparsity of the available datasets, so that no biological information may be available for the end-points of a link in the PPI. In this case, the link receives a minimum weight c. As with personalization vectors, gene expression information about a population of patients can also be used to define a tissue-specific, population-dependent transition matrix. Specifically, gene expression information is used to assign weights to edges of the underlying PPI network, this time reflecting similarities between genes in terms of co-expression with respect to the subject population. Consider a CO-Expression network in which each pair of genes (i, j) is assigned a score equal to the Pearson's correlation coefficient $pc_{i,j}$ between the expression levels of i and j within the population. We can define a transition matrix W by assigning each edge (i, j) of the PPI network a probability as follows:

$$W_{ij} = \frac{|pc_{ij}|}{\sum_{k \in N(i)} |pc_{ik}|} \tag{6}$$

where $N(i)$ is the set of i's neighbors in the PPI network. Note that (i) for every node/gene i we consistently have a probability distribution over its incident edges and (ii) the importance of edges reflects the absolute value of the correlation between the expression levels of two genes within the population of interest.

Integrating biological information and gene expression. We discussed above two orthogonal approaches to the design of personalization vectors. The first one leverages similarities between the biological processes associated to known disease proteins and those to be prioritized. Hence, teleporting probabilities depend on the seed set through association with common biological processes. In the second case, teleporting probabilities depend on information that is tissue-specific (the level of expression in a population of subjects affected by a certain disease) and partly on the seed set, but this time through the PPI's topology. Hence, these two approaches largely rely on complementary sources of information. In order to integrate these complementary sources into a unique personalization vector that leverages both, we follow a simple, yet mathematically principled approach, whereby we take a *convex combination* of the corresponding personalization vectors. Namely, assume, we have computed two personalization vectors $\mathbf{q}_1$ and $\mathbf{q}_2$, the former using biological information only, the latter using gene expression data and the PPI. We obtain a personalization vector as follows:

$$\mathbf{q} = \alpha \mathbf{q}_1 + (1 - \alpha)\mathbf{q}_2, \tag{7}$$

where $\alpha \in [0, 1]$. Parameter α allows to weigh in the relative importance of the different information sources we are using. Intuitively,

this type of aggregation amounts to considering a gene a potential candidate if it is statistically significant in terms of its involvement in biological processes, of its gene expression levels in the subject group, or both (note that this approach seamlessly extends to an arbitrary number of information sources). The choice of α (and other parameters of the model), its impact on performance and dependence of the optimal choice on the scenario at hand are discussed in detail in Sections 3 and 4 and Supplementary Section C.2. Remark. Despite its simplicity, this is a mathematically principled choice. In particular, it is well-known (Jeh and Widom, 2003) and easy to show that the stationary distribution corresponding to the convex combination of two personalization vectors $\mathbf{q}_1$ and $\mathbf{q}_2$ is itself the linear combination of the stationary distributions corresponding to $\mathbf{q}_1$ and $\mathbf{q}_2$, respectively. Briefly put, the parameter α allows us to tune the relative importance of the information provided by gene annotations and gene expression levels, respectively. Moreover, this approach extends seamlessly to an arbitrary number of heterogeneous sources of information (and corresponding personalization vectors). Other aggregation methods were also considered, yet they provided worse or at most comparable results. Some are presented in Supplementary Section B.3.2 for the sake of completeness. In the previous paragraphs, we have seen how we can derive random walk transition matrices using only information about biological processes (i.e. annotations) or gene expression data from a population of subjects. As with personalization vectors, multiple transition matrices derived from complementary biological sources can be integrated into a single *aggregate transition matrix*. For example, assume, we computed transition matrices W_1 and W_2 using biological and gene expression information, respectively. Any convex combination of W_1 and W_2 is a feasible transition matrix. Namely, for some $\beta \in [0, 1]$, we consider the transition matrix

$$W = \beta W_1 + (1 - \beta) W_2.$$

It is easy to see that (i) W is still a transition matrix, i.e. the entries of each row sum to one and that (ii) the approach seamlessly extends to any number of complementary sources. In the case of biological annotations and gene expression data, $\beta = 1$ corresponds to only considering biological annotations, whereas $\beta = 0$ corresponds to only leveraging gene expression data. So, β is a parameter, whose tuning allows us to weigh the importance of one source of information with respect to the other.

3 Results

This section investigates BRW's performance in prioritizing gene candidates for genetic diseases.

3.1 Experimental setup

We used a number of biological data sources. Some of them were used as inputs, to define key parameters of our algorithms, while others were used to biologically validate the results of the algorithms we considered. They are briefly described here and more extensively in Supplementary Section A.

Data sources. The experiments discussed in Section 3 were conducted on the HIPPIE PPI network (Alanis-Lobato *et al.*, 2016) and on the same PPI network as in Ghiassian *et al.* (2015) for the sake of comparison.

We used three different sources of gene biological information: GO Consortium (http://geneontology.org/) where, for each gene, we downloaded its biological processes, KEGG (Kanehisa, 2019; Kanehisa and Goto, 2000; Kanehisa *et al.*, 2019) and Reactome (Jassal *et al.*, 2020), which we used to download pathways' annotations. Gene expression datasets were downloaded from The Cancer Genome Atlas database (https://portal.gdc.cancer.gov/).

We validated the methods considered in this study both via an internal validation on known disease genes and through an external validation using drug target associations. In more detail, we used disease–gene associations as in Ghiassian *et al.* (2015) that describe a corpus of 70 Mendelian diseases. From Piñero *et al.* (2020), we further derived known disease–gene associations for the four different cancer types that we investigate in Section 3 (i.e. breast cancer, lung adenocarcinoma, papillary thyroid cancer and colorectal adenocarcinoma). Finally, we used Drug–Gene Target associations from DrugBank (https://go.drugbank.com/) (Wishart *et al.*, 2017). We selected only drugs approved for breast cancer treatment from FDA (https://www.cancer.gov/about-cancer/treatment/drugs/breast) (Supplementary Table S2).

All data sources mentioned above are more extensively discussed in Supplementary Section A.

Baselines. To provide a robust performance assessment, we compared BRW with a number of well-known, state-of-art baselines for disease gene prioritization, namely, RWR (Köhler *et al.*, 2008), DIAMOnD (Ghiassian *et al.*, 2015), DADA (Erten *et al.*, 2011) and RWR-M (Valdeolivas *et al.*, 2019). For the sake of completeness, Supplementary Section C.4 also compares our approach against an embryonic (and underperforming) version of our framework that was presented in Gentili *et al.* (2019).

Performance indices. We used the widely adopted indices Recall@K and nDCG in some of the experiments described in the remainder. To keep presentation self-contained, they are briefly described in Supplementary Section C.1.

3.2 Multi-omics integration improves algorithm performance

In a first round of experiments, we performed two (internal) validation steps: (i) we first compared the algorithms with respect to four cancer phenotypes (i.e. breast cancer, lung adenocarcinoma, papillary thyroid cancer and colorectal adenocarcinoma), for which both biological annotations and gene expression data are available; and (ii) as a further step, we performed a broader, yet less specific validation on a corpus of 70 manually curated Mendelian diseases (Ghiassian *et al.*, 2015), for which only biological annotations were used. For each disease, we performed a mean 100-fold validation, by sampling 70% of known disease genes uniformly at random and using the rest to test the algorithms. For each tested disorder, we computed Recall@K and the nDCG of each algorithm.

Four cancer phenotypes. We performed a grid search to identify the best combination of hyper-parameters (α, β, r). The benchmark, discussed in Supplementary Section C.2 and illustrated in Supplementary Figures S6 and S7, highlights interesting, mixed results. For the task of disease gene prioritization [we remark that the settings discussed here are not optimal for other tasks in general, e.g. drug target discovery (see Section 3.4)] and to the purpose of computing personalization vectors (i.e. teleporting probabilities), the signal contained in statistically significant annotations derived from seed genes is definitely stronger than the signal carried by gene expression, so that the best choice for disease gene prioritization on cancer phenotypes is $\alpha = 1.0$, thus, completely removing information about DE genes. However, gene expression information is crucial in determining the transition matrix of the random walk, with values of $\beta \in [0.25, 0.5]$ achieving best predictive performance, indicating that both information sources provide crucial information to BRW for disease gene ranking.

Results for the four cancer phenotypes show that algorithms that only leverage the PPI tend to perform worse in terms of both Recall@k and nDCG, as shown in Figure 2a and b. Conversely, multi-Omics methods that, like BRW and RWR-M, rely on multiple biological information sources typically perform better in terms of the aforementioned indices. Improvement of these methods over single-source baselines at least in part stems from the well-known fact that disease-associated genes tend to be involved in similar pathways and biological processes (Barabási *et al.*, 2011), see also Section 3.3. Interestingly, BRW's ability to bias the random walk toward related genes that do not necessary belong to the seed set seems to play an important role in improving prioritization of the test set, at least in terms of Recall@k. At the same time, RWR-M achieves similar performance (slightly better or worse, depending on the dataset) if one considers nDCG as a global measure of rank, as shown in Figure 2b. As previously remarked, biased teleporting makes BRW explore areas of the PPI network that could be

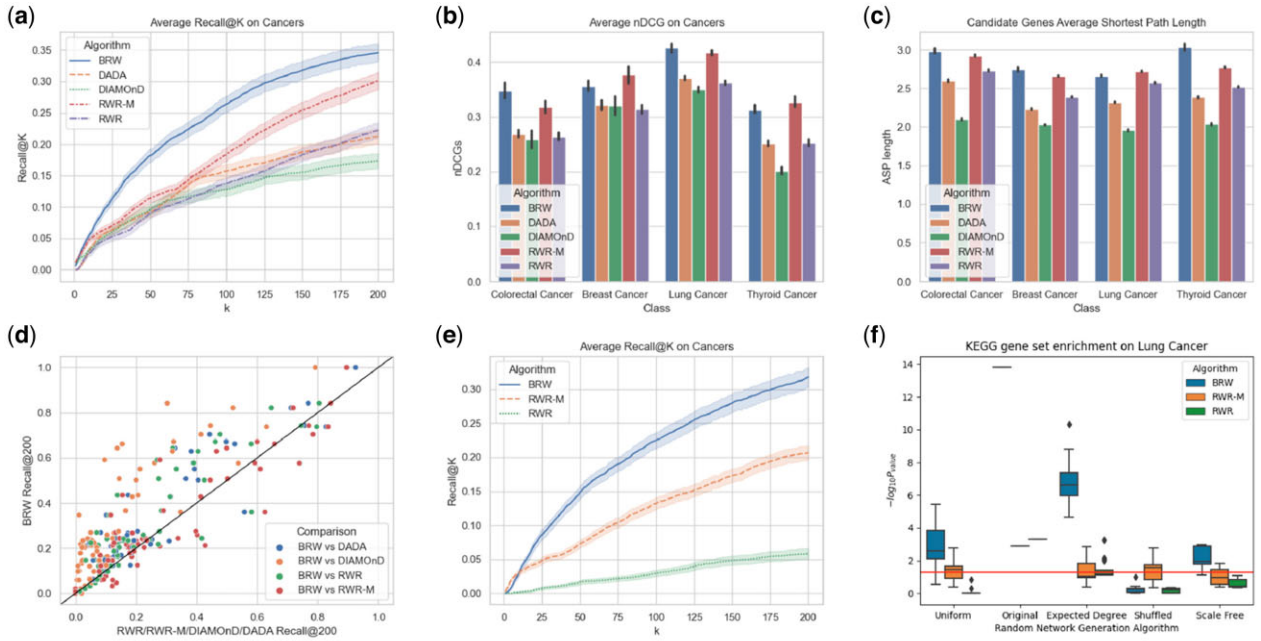

Fig. 2. Algorithm comparison: (**a**) and (**b**) compare, respectively, the Recall@K and nDCG of some of well-known network-based approaches with BRW on four types of cancer using a 100-fold Monte Carlo random sampling validation choosing uniformly at random the 70% of known disease genes and considering the remaining ones as the test set (30%). (**c**) Shows the average shortest path length of top 200 candidates predicted by the analyzed algorithms. (**d**) Shows the comparison between BRW with the state-of-the art on a corpus of 70 Mendelian diseases downloaded from Ghiassian *et al.* (2015). (**e**) Compares multi-omics frameworks (BRW and RWR-M) on average Recall@k on cancer phenotypes when the PPI network is randomized. The average Recall@K is computed as described in the previous experiment. (**f**) Illustrates how multi-omics integration affects the bias induced by curated ontologies, such as GO, Kegg and Reactome and the PPI network. It compares the distribution of *P*-values (negative log scale) computed by the testsuit using KEGG pathways and gseapy on gene sets predicted on randomized PPI networks with those predicted on the original PPI. As hypothesized, BRW, having in input statistical significant pathways from KEGG, Reactome and GO, returns biologically meaningful candidate sets. However, their significance is not comparable with the set predicted using the original PPI

relatively far from the seed set, a fact that is reflected in its candidate genes in the top 200 positions having higher average shortest path distance than other RWR-based methods that only teleport to genes of the seed set, as shown in Figure 2c.

Mendelian diseases. As a further internal validation, we provided a less specific yet broader, comparative assessment of BRW, by performing a Monte Carlo cross-validation (Dubitzky *et al.*, 2007) on 'gold standard' gene sets. These sets contain known genes associated with 70 diseases, which were previously selected in Ghiassian *et al.* (2015) from OMIM and PheGenI databases. As discussed more in detail at the end of Supplementary Section A.1, in this experiment (and only in this one), we used the PPI considered in Ghiassian *et al.* (2015), for the sake of consistency [all other experiments use the more recent HIPPIE-v2.2 PPI network of Alanis-Lobato *et al.* (2016)]. In this second round of experiments, we set $\alpha = 1.0$ and $\beta = 1.0$, since gene expression is not used. Figure 2d compares the performances of RwR, DIAMOnD, DADA, RWR-M and BRW in terms of Recall@k. Our heuristic ranks more known disease proteins than DIAMOnD in the top 200 positions for 95% of disorders analyzed and respectively for 72%, 73% and 63% of the disorders for RwR, DaDA and RWR-M.

3.3 Randomization and bias

BRW and RWR-M leverage multiple data sources. As a result, we expect their results to be less affected by random noise in the PPI with respect to other heuristics. To test this hypothesis, we performed a first experiment, by performing an internal validation as done in Section 3.2, but this time running the random walk-based heuristics using degree-preserving randomized version of the PPI. To this purpose, we implemented the degree-preserving randomization algorithm of Milo *et al.* (2003), also described in Supplementary Section B.5. As hypothesized (Fig. 3e), by leveraging multiple biological sources, BRW and RWR-M are less affected by randomization of the PPI. This effect is stronger in BRW, whose teleporting probabilities also depend on phenotype information (statistically significant ontologies, pathways and DE genes). This is further shown in the more detailed Supplementary Figure S8a, highlighting a positive correlation between the number of test genes ranked in the first 200 positions (Recall@K) and the value of BRW's restart probability *r*. As remarked above, this effect is also present in RWR-M, in which heterogeneous biological sources are summarized in different, layered networks. In particular, as shown in Supplementary Figure S8a, RWR-M's performance degrades significantly if one also randomizes the Pathways network used by the algorithm. Altogether, these results suggest that phenotypical information associated to the nodes plays an important role in prioritization for these heuristics, owing to the fact that known disease proteins tend to be involved in the same pathways, a fact that becomes apparent in an internal-only validations as the one considered in this section.

The results above are part of a more general phenomenon. As shown in Lazareva *et al.* (2021), several heuristics inherit biases present in up-to-date PPI networks and manually curated ontologies, such as GO, KEGG and Reactome, in some cases with results that improve when the PPI is replaced by a randomized one. To further investigate this issue, in particular, the bias inherited by BRW when ontologies from manually curated data sources are used, we considered the Active Module Identification Methods test suite proposed in Lazareva *et al.* (2021), which allows systematic comparison of candidate gene sets predicted using an original PPI network and perturbed versions thereof. To this purpose, we considered the Lung Cancer phenotype, which was both considered in Section 3.2 and is used as a benchmark in Lazareva *et al.* (2021). We considered randomized versions of the original PPI obtained using all randomization algorithms implemented in the test suite (i.e. expected degree, uniform, shuffled and scale-free). Candidate genes ranked by BRW, RWR-M and RWR using the original PPI as input were compared with the results obtained using each of the aforementioned randomized counterparts of the PPI. Statistical significance was

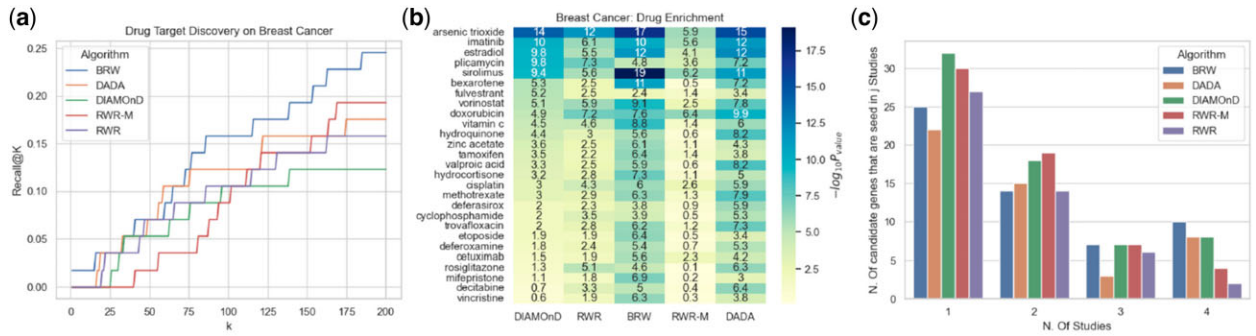

Fig. 3. Algorithm comparison on breast cancer disease: (**a**) percentage of breast cancer drug targets found by each framework in the top K positions. (**b**) Drugs that are enriched (corrected P-value $< 10^{-5}$) in at least one of the predicted candidate gene sets. (**c**) Number of candidate genes, predicted by each algorithm, that are frequently mutated in at least j of the remaining studies, with $j \in \{1, 2, 3, 4\}$

computed on KEGG pathways enrichments computed by the test suite on predefined KEGG pathway.

Figure 2f shows that BRW, RWR-M and RWR extract statistically significant candidate gene sets on the original PPI network. Relying only on the PPI, RWR provides no statistically significant results when the PPI is randomized, except when expected degree is preserved. (This phenomenon arises because the stationary distribution of a random walk on an undirected network follows exactly the degree distribution. As a result, the stationary distribution of a RWR over an undirected network is positively correlated with degree distribution.) On the other hand, BRW and RWR-M inherit PPI and ontology's biases, so that their results are still statistically significant when the PPI is randomized using expected degree, scale-free and uniform randomization algorithms. Still, BRW's results are considerably more significant when the original network is used, indicating that topology of the PPI is crucial to (more) effectively propagate information extracted from other biological sources. BRW predicts more significant outcomes on the actual PPI than on the randomized versions. While all ontologies (GO, KEGG and Reactome) were considered in the experiment summarized in Figure 2f, we further investigated BRW's behavior when KEGG (which is used in the enrichment) is not used. In this case, BRW's results are no longer statistically significant if one randomizes the network, with the exception of a mild significance when expected degree is preserved.

3.4 A case study: breast cancer phenotype

In this section, we discuss the results of an in-depth analysis of BRW's performance on the breast cancer phenotype, a global health concern (Crimini *et al.*, 2021; Siegal *et al.*, 2014), with 284 200 new cases and more than 44 000 deaths in the USA in 2021 (Siegal *et al.*, 2014). In particular, we present (i) an external validation using drugs FDA-approved drugs; (ii) drug enrichment; and (iii) an assessment of the algorithm(s) stability across multiple populations.

3.4.1 Drug target discovery

We validated the top candidate gene sets prioritized by each algorithm along different axes. To this purpose, we created a test set of target genes for breast cancer drugs approved by the FDA as described in Section 3.1. Results are summarized in Figure 3a, showing Recall@K for the algorithms we considered. Compared to other baselines, BRW prioritizes the highest number (25%) of drug targets in the top 200 candidates. Looking at the specific gene targets predicted by the algorithms, BRW predicts the highest number of genes (14 out of the 20 predicted genes), making four unique predictions, namely, Cyclin Dependent Kinases 4 and 6 (*CDK4* and *CDK6*), the Protein Kinase C Zeta (*PRKCZ*) and Caspase 3 (*CASP3*). On the other hand, the DNA Topoisomerase II Alpha (*TOP2A*) and the progesterone receptor (*PGR*) genes are uniquely predicted by DIAMOnD. The Protein Kinase C Theta gene (*PRKCQ*) is only predicted by RWR-M algorithm. More details are given in Supplementary Table S5.

Furthermore, we validated the prioritized genes from a different perspective: we considered the group of drug targets returned by each framework and showed the drugs that target them. We filtered out drugs in Drug Bank that were not annotated with the 'Breast Cancer' or related associated condition. Supplementary Table S3 shows how the number of drugs ranges from 6 to 17 as we change the combination of hyper-parameters. As expected, the number of drugs correlates positively with the predicted drug target percentage (Recall@200). Supplementary Table S4 shows the drugs prioritized (i.e. a drug with at least one drug target prioritized in the top 200 positions) by BRW and the other algorithms. BRW prioritizes the highest number of drugs. Interestingly, CDK4/6 inhibitors palbociclib, ribociclib and abemaciclib, currently used to treat hormone receptor-positive/HER2-negative metastatic breast cancer (Duranti *et al.*, 2021), are only predicted by BRW.

Finally, we identified drugs that have an enrichment with the top candidate genes predicted by each algorithm. To identify enriched drugs, we used the gseapy package for gene set enrichment analysis (Subramanian *et al.*, 2005), and we chose drugs that were enriched by at least one algorithm with a corrected P-value lower than 10^{-5}. In particular, Supplementary Table S3 shows how the enrichment is affected by various hyper-parameters combinations. In this case, drug enrichment correlates positively with gene expression, and the best combination is obtained when α and β are equal to 0.25. Indeed, while drugs target and inhibit genes involved in disease-specific pathways, the effect of the drugs can be measured by the differential expression and the co-expression between targets and nearby genes (Chen *et al.*, 2017). In general, results highlight the following trends: (i) different biological sources (in our case, gene expression and ontologies) provide complementary information, with different subsets of FDA-approved drug targets significantly enriched for both low (gene expression bias) and high (ontologies bias) values of the parameters α and β; and (ii) best performance is achieved for lower values of the restart probability r (0.25), confirming that BRW's prioritization depends on information that is not necessarily confined to the immediate neighborhood of the seed set.

Figure 3b highlights drugs that are enriched in BRW and in the other baselines considered in this manuscript, with details for each drug reported in Supplementary Table S6. Interestingly, 11 FDA-approved drugs for breast cancer treatment are enriched in BRW gene candidates (i.e. fulvestrant, doxorubicin, paclitaxel, tamoxifen, methotrexate, letrozole, cyclophosphamide, trastuzumab, fluorouracil and gemcitabine). The mTOR inhibitor sirolimus showed the best significance. While it is not currently used to treat breast cancer, preclinical *in vivo* studies demonstrated its potent antiangiogenic activity on breast cancer models (Muhammad Sakri *et al.*, 2022).

3.4.2 Breast cancer—multi-population study

A desirable property of disease gene prioritization should be a certain stability in the set of proposed disease gene candidates across different populations. In other words, to some extent (e.g. up to intrinsic biases or qualitative differences in the datasets used), results

should not overly depend on the specific dataset the algorithm is analyzing. To investigate this aspect, we characterized the behaviors of BRW and the other baselines we considered when applied to different population studies on breast cancer. We considered Invasive Breast Cancer population studies retrieved from cBioPortal datasets (Banerji *et al.*, 2012; Ciriello *et al.*, 2015; Kan *et al.*, 2018; Shah *et al.*, 2012; Stephens *et al.*, 2012) (see Supplementary Section A), which allowed us to identify five different seed sets for the algorithms we considered, one per study. Figure 3d plots the number of candidate genes, predicted by each algorithm, that are seeds in at least j other studies (frequently mutated genes in the associated populations, i.e. mutation frequency >1%), for $j \in \{1, 2, 3, 4\}$. Notably, BRW has the highest number of candidates that are frequently mutated in the other four studies (the far right bar). The 10 genes identified by BRW, that are frequently mutated in four out of the five breast cancer populations are *AR*, *JUN*, *STAT3*, *NOTHC1*, *JAK2*, histone deacetylase 1 (*HDAC1*), *SMAD4*, *YAP1* and *CHD4*. While the *HDAC1* is also retrieved by all the other algorithms, *SMAD4* and *CHD4* are only predicted by BRW. The remaining genes are returned by BRW and at least another heuristic. Details are reported in Supplementary Table S7.

4 Discussion

Guided by the hypothesis that disease causing genes often share important common biological processes and pathways, we extended the RWR approach to disease gene prioritization, proposing a framework that allows seamless integration of multiple biological information sources. The proposed approach consists of two main steps: (i) extracting significant disease features from disease-term association data, such as statistically significant biological processes and pathways, and (ii) using these features to bias the RWR in a way that is consistent with the biological sources used. These two aspects are discussed in Section 2.

In general, BRW outperforms, in terms of standard indices of predictive accuracy, such as recall and nDCG, previous frameworks that only rely on a s single biological source, such as RWR (Navlakha and Kingsford, 2010), DaDa (Erten *et al.*, 2011) and DIAMOnD (Ghiassian *et al.*, 2015) that rely on PPI network. This is also true for an extension of RWR, namely, RWR-M (Valdeolivas *et al.*, 2019), that performs a random walk on a multi-layer network. Using a Monte Carlo random sampling validation, we showed that prioritization results returned by BRW frequently outperform other baselines on four different cancer types: breast, colorectal, lung, and thyroid cancer. These results were further supported by a broader, computational validation on a corpus of 70 Mendelian disease manually curated by Ghiassian *et al.* (2015).

A main aim of precision medicine is to use disease genes to enable tailored treatments. In this perspective, we investigated how the candidate genes prioritized by each framework are related to breast cancer drug targets. Results show that BRW prioritizes the highest number of drug targets in the top 200 candidates (Fig. 3a). As a further assessment, we considered the group of drug targets returned by each algorithm and identified the drugs that target them. We found that BRW prioritizes the highest number of FDA-approved drugs for breast cancer treatment (Supplementary Table S4). In general, we noticed that drug targets are more correlated with phenotypic pathways, while drug enrichment highlights how drugs affect gene expression in terms of co-expression and differential expression. This is not surprising: indeed, although drugs often target and inhibit genes involved in disease-specific pathways, the effect of the drugs can be measured by the differential expression and the co-expression between targets and nearby genes (Chen *et al.*, 2017).

To investigate the stability of the proposed disease gene candidates, we characterized the behaviors of BRW and the previous frameworks across different population studies. We selected 'invasive breast cancer' as a phenotype and we retrieved data for five different populations from cBioPortal (https://www.cbioportal.org/). Consistently, BRW showed excellent stability, with the highest number of gene candidates in one study that are frequently mutated in the other four (Fig. 3c and Supplementary Table S7).

The BRW framework is not without limitations. On one hand, inducing a bias in teleporting probability and the transition matrix through the use of ontologies improves predictive accuracy of the algorithm. On the other hand, this bias can hinder BRW's ability to identify new disease-related pathways. For this reason, we believe it is important to exploit the framework's ability to integrate heterogeneous, hopefully complementary, data sources. Furthermore, experiments summarized in Supplementary Figure S7 and Supplementary Table S3 quantitatively showed that every data source comes with its bias. As a result, the relative weights attributed to different biological information sources (the combined choice of α and β in our case) can significantly affect predictive accuracy, with a magnitude that in general depends on the validation test used. For example, while ontologies showed most effective in a computational validation and for drug target discovery, gene expression proved particularly useful in identifying a candidate gene set that is highly enriched in breast cancer and cancer related drugs. In our opinion, these results provide support to the use of (at least partially) complementary sources of biological information, even though these sometimes present non-negligible correlations, as discussed in Section 3.3.

We also emphasize that, in this study, we only leveraged a limited set of disease-specific data sources (co-expression and differential expression). Potential performance improvements might be achieved by integrating further disease information sources, such as methylation data, microRNA expression or microRNA-target gene associations. Doing this might provide further insights into key biological mechanisms and provide new prospective gene targets, though, of course, only functional studies can provide the ultimate answer as to their biological role.

Funding

This work was partially supported by the European Research Council (ERC) Advanced Grant 788893 AMDROMA 'Algorithmic and Mechanism Design Research in Online Markets'; the European Commission (EC) H2020RIA project 'SoBigData++' (871042); and the Ministero Istruzione, Università e Ricerca (MIUR) Progetto di Rilevante Interessa Nazionale (PRIN) project ALGADIMAR 'Algorithms, Games, and Digital Markets'.

Conflict of Interest: none declared.

References

Alanis-Lobato,G. *et al.* (2016) HIPPIE v2. 0: enhancing meaningfulness and reliability of protein–protein interaction networks. *Nucleic Acids Res.*, **45**, D408–D414.

Banerji,S. *et al.* (2012) Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature*, **486**, 405–409.

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

Chen,B. *et al.* (2017) Reversal of cancer gene expression correlates with drug efficacy and reveals therapeutic targets. *Nat. Commun.*, **8**, 16022.

Ciriello,G. *et al.*; TCGA Research Network. (2015) Comprehensive molecular portraits of invasive lobular breast cancer. *Cell*, **163**, 506–519.

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

Crimini,E. *et al.* (2021) Precision medicine in breast cancer: from clinical trials to clinical practice. *Cancer Treat. Rev.*, **98**, 102223.

De Bie,T. *et al.* (2007) Kernel-based data fusion for gene prioritization. *Bioinformatics*, **23**, i125–i132.

Dimitrakopoulos,C. *et al.* (2018) Network-based integration of multi-omics data for prioritizing cancer genes. *Bioinformatics*, **34**, 2441–2448.

Dubitzky,W. *et al.* (2007) *Fundamentals of Data Mining in Genomics and Proteomics*. Springer Science & Business Media, New York.

Duranti,S. *et al.* (2021) Breast cancer drug approvals issued by EMA: a review of clinical trials. *Cancers*, **13**, 5198.

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

Gentili,M. *et al.* (2019) Biological random walks: integrating heterogeneous data in disease gene prioritization. In: *2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. pp. 1–8. IEEE.

Ghiassian,S.D. *et al.* (2015) A disease module detection (diamond) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. *PLoS Comput. Biol.*, **11**, e1004120.
Gustafsson,M. *et al.* (2014) Modules, networks and systems medicine for understanding disease and aiding diagnosis. *Genome Med.*, **6**, 82.
Hardy,J. and Singleton,A. (2009) Genomewide association studies and human disease. *N Engl. J. Med.*, **360**, 1759–1768.
Jassal,B. *et al.* (2020) The reactome pathway knowledgebase. *Nucleic Acids Res.*, **48**, D498–D503.
Jeh,G. and Widom,J. (2003) Scaling personalized web search. In: *Proceedings of the 12th International Conference on World Wide Web*. pp. 271–279.
Kan,Z. *et al.* (2018) Multi-omics profiling of younger Asian breast cancers reveals distinctive molecular signatures. *Nat. Commun.*, **9**, 1–13.
Kanehisa,M. (2019) Toward understanding the origin and evolution of cellular organisms. *Protein Sci.*, **28**, 1947–1951.
Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
Kanehisa,M. *et al.* (2019) New approach for understanding genome variations in KEGG. *Nucleic Acids Res.*, **47**, D590–D595.
Koboldt,D. *et al.* (2013) The next-generation sequencing revolution and its impact on genomics. *Cell*, **155**, 27–38.
Köhler,S. *et al.* (2008) Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.*, **82**, 949–958.
Korcsmaros,T. *et al.* (2017) Next generation of network medicine: interdisciplinary signaling approaches. *Integr. Biol. (Camb)*, **9**, 97–108.
Lazareva,O. *et al.* (2021) On the limits of active module identification. *Brief. Bioinform.*, **22**, bbab066.
Levin,D.A. and Peres,Y. (2017) *Markov Chains and Mixing Times*. Vol. **107**. American Mathematical Soc, Providence, Rhode Island, USA.
Menche,J. *et al.* (2015) Uncovering disease-disease relationships through the incomplete interactome. *Science*, **347**, 1257601.
Menche,J. *et al.* (2017) Integrating personalized gene expression profiles into predictive disease-associated gene pools. *NPJ Syst. Biol. Appl.*, **3**, 10.
Milo,R. *et al.* (2003) On the uniform generation of random graphs with prescribed degree sequences. *arXiv preprint cond-mat/0312028*.
Muhammad Sakri,M.S. *et al.* (2022) Rapamycin as a potent and selective inhibitor of vascular endothelial growth factor receptor in breast carcinoma. *Int. J. Immunopathol. Pharmacol.*, **36**, 20587384211059673.
Navlakha,S. and Kingsford,C. (2010) The power of protein interaction networks for associating genes with diseases. *Bioinformatics*, **26**, 1057–1063.
Ozturk,K. *et al.* (2018) The emerging potential for network analysis to inform precision cancer medicine. *J. Mol. Biol.*, **430**, 2875–2899.
Piñero,J. *et al.* (2020) The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.*, **48**, D845–D855.
Shah,S.P. *et al.* (2012) The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature*, **486**, 395–399.
Shang,H. and Liu,Z.-P. (2020) Network-based prioritization of cancer genes by integrative ranks from multi-omics data. *Comput. Biol. Med.*, **119**, 103692.
Siegal,R. *et al.* (2014) Cancer statistics, 2012. *CA Cancer J. Clin.*, **64**, 9–29.
Silverman,E.K. *et al.* (2020) Molecular networks in network medicine: development and applications. *Wiley Interdiscip. Rev. Syst. Biol. Med.*, **12**, e1489.
Stephens,P. *et al.*; Oslo Breast Cancer Consortium (OSBREAC). (2012) The landscape of cancer genes and mutational processes in breast cancer. *Nature*, **486**, 400–404.
Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA*, **102**, 15545–15550.
The Gene Ontology Consortium. (2019) The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.*, **47**, D330–D338.
Valdeolivas,A. *et al.* (2019) Random walk with restart on multiplex and heterogeneous biological networks. *Bioinformatics*, **35**, 497–505.
Wishart,D. *et al.* (2017) DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.*, **46**, D1074–D1082.