

Contextualizing and interpreting biological data with pathway knowledge and network algorithms

Kumulative Dissertation

zur Erlangung des Doktorgrades (Dr. rer. nat.)
der Mathematisch-Naturwissenschaftlichen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

SARAH MUBEEN

aus Morgantown, United States of America

Bonn, 2022

Angefertigt mit Genehmigung
der Mathematisch-Naturwissenschaftlichen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Gutachter: Univ.-Prof. Dr. rer. nat. Martin
Hofmann-Apitius
2. Gutachter: Univ.-Prof. Dr. rer. nat. Diana Imhof

Tag der Promotion: 23rd January 2023
Erscheinungsjahr: 2023

Abstract

Elucidating the mechanisms which give rise to unique biological functions is a major goal of human biology. Vast quantities of biological data that have been amassed can aid us in understanding these mechanisms, such as at the level of gene expression, protein expression and metabolites, although piecing together these various components can be an arduous undertaking. Though it may be difficult to ascertain the interplay of these entities, living organisms are complex and composed of highly interconnected and interdependent systems, thus requiring system-wide investigations. To that end, the complex interplay of biological entities can be modelled in interaction networks on which various computational approaches can subsequently be applied. In this work, we augment existing network-based approaches and introduce novel ones towards the aim of building a more comprehensive picture of the mechanisms that regulate health and disease. Firstly, we address this goal by building upon existing techniques for the interpretation of high dimensional data through its representation as biological pathways. We integrate pathway knowledge dispersed across heterogeneous resources for a more comprehensive overview of the current knowledge surrounding a particular biological process under investigation. We also outline how various aspects of widely-used pathway analyses contribute to either sound or misinterpretations. Secondly, we combine a knowledge and data-driven approach in order to contextualize gene expression data and reveal transcriptional patterns underlying discrete biological functions across four distinct contexts (i.e., disease, tissue, cell type and cell line). Thirdly, we introduce novel network-based methodologies for biomedical applications. These include implementations of network diffusion algorithms alongside of several multimodal biological networks to operate on, as well as a pathfinding algorithm for drug discovery. In conclusion, compiling biological data into networks of interacting molecular entities and biological constructs enables us to achieve ever-increasing levels of discernment of the mechanisms that govern health and disease.

Know the truth. Do the good. Make the beautiful.

The transcendentals

Acknowledgments

I truly feel fortunate to have the opportunities I did in the department of Bioinformatics under the guidance of Prof. Dr. Martin Hofmann-Apitius. Thank you for all the support you've given me over the past few years, both academic and personal. I would also like to express my gratitude to Prof. Dr. Diana Imhof for acceding to be an additional reviewer of my thesis as well as to the members of my defence committee.

To Dr. Daniel Domingo-Fernández, my academic advisor and scientific mentor, I really can't seem to find the right words to say thank you. You made all of this possible. For the wealth of knowledge and all the guidance you've given me the past five years and for the scientist you've shaped me into, I owe you my sincerest gratitude. Your foresight and determination were the driving force for this work and you taught me such a great deal about perseverance. You taught me technical skills when I had none and to see the big picture when I was stuck in the details. Your expertise and your cleverness have taken me by surprise on many an occasion (although admittedly, my surprise could only ever be attributed to your humility) and helped me to navigate through challenges which I had thought were impassable. Ultimately, I know that you're a wonderful scientist and a brilliant academic, but more than that, I know that you're an even better person who follows his dreams to the fullest and encourages each person he encounters to do the same. Thanks coach!

Special thanks are also in order for all of my fellow PhD students and colleagues at SCAI. Though the duration of my PhD largely coincided with the COVID-19 pandemic and our time together was ever so short, I am so grateful for all the wonderful discussions and engagement that still took place. In particular, to Rebeca Queiroz Figueiredo, Vinay Srinivas Bharadhwaj and Sepehr Golriz Khatami, it was such a pleasure working with you! To Alpha

Tom Kodamullil, I am so glad to have had your support all these years since I first started at SCAI over six years ago. A heartfelt thank you to Meike Knieps and Alina Enns for always keeping things running so smoothly and for all of your patience. Warm regards to the Women in Science! It was my good fortune to have had the constant support of such talented and kind colleagues like you!

Finally, to my family and friends who are never so far, cheering me on and helping me to keep things in perspective, I am so thankful. To my mom and dad, my appreciation and gratitude towards you are boundless. You instilled in me the importance of integrity and a commitment to always seek knowledge, lessons that have stayed with me throughout my academic journey.

Declaration

I hereby certify that this material is my own work, that I used only those sources and resources referred to in the thesis, and that I have identified citations as such.

Sarah Mubeen

Publications

Thesis Publications

† Joint last authors

1. **Mubeen S.**, Kodamullil A.T., Hofmann-Apitius M., and Domingo-Fernández D. (2022). On the influence of several factors on pathway enrichment analysis. *Briefings in Bioinformatics*, 23(3).
<https://doi.org/10.1093/bib/bbac143>
2. **Mubeen S.**, Hoyt C. T., Gemünd A., Hofmann-Apitius M., Fröhlich H., and Domingo-Fernández D. (2019). The Impact of Pathway Database Choice on Statistical Enrichment Analysis and Predictive Modeling. *Frontiers in Genetics*, 10:1203.
<https://doi.org/10.3389/fgene.2019.01203>
3. **Mubeen S.**, Bharadhwaj S. V., Kodamullil A.T., Gadiya Y., Hofmann-Apitius M., and Domingo-Fernández D. (2021). DecoPath: A Web Application for Decoding Pathway Enrichment Analysis. *NAR Genomics and Bioinformatics*, 3(3): lqab087.
<https://doi.org/10.1093/nargab/lqab087>
4. Figueiredo R.Q.,[†] Raschka T.,[†] Kodamullil A.T., Hofmann-Apitius M., **Mubeen S.**[†], and Domingo-Fernández D.[†] (2021). Towards a global investigation of transcriptomic signatures through co-expression networks and pathway knowledge for the identification of disease mechanisms. *Nucleic acid research*, 49(14): 7939–7953.
<https://doi.org/10.1093/nar/gkab556>

5. Figueiredo R.Q., Díaz del Ser S., Raschka T., Hofmann-Apitius M., Kodamullil A.T., **Mubeen S.**[†], and Domingo-Fernández D.[†] (2022). Elucidating gene expression patterns across multiple biological contexts through a large-scale investigation of transcriptomic datasets. *BMC Bioinformatics*, 23(1).
<https://doi.org/10.1186/s12859-022-04765-0>
6. Marín-Llaó J., **Mubeen S.**, Perera-Lluna A., Hofmann-Apitius M., Picart-Armada S.[†], and Domingo-Fernández D.[†] (2021). MultiPaths: a Python framework for analyzing biological networks using diffusion algorithms. *Bioinformatics*, 37(1): 137-139.
<https://doi.org/10.1093/bioinformatics/btaa1069>
7. Rivas-Barragan D., **Mubeen S.**, Guim Bernat F., Hofmann-Apitius M., and Domingo-Fernández D. (2020). Drug2ways: Reasoning over causal paths in biological networks for drug discovery. *PLOS Computational Biology*, 16(12): e1008464.
<https://doi.org/10.1371/journal.pcbi.1008464>

Other Publications

8. Domingo-Fernández D., **Mubeen S.**, Marín-Llaó J, Hoyt C. T., and Hofmann-Apitius M. (2019). PathMe: Merging and exploring mechanistic pathway knowledge. *BMC Bioinformatics*, 20(1): 243.
<https://doi.org/10.1186/s12859-019-2863-9>
9. Hoyt C. T., Domingo-Fernández D., **Mubeen S.**, Marín-Llaó J, Kono-topetz A., Ebeling C., ... Hofmann-Apitius M. (2019). Integration of Structured Knowledge Sources with Biological Expression Language. *bioRxiv*, 631812.
<https://doi.org/10.1101/631812>
10. Golriz Khatami S., **Mubeen S.**, and Hofmann-Apitius, M. (2020). Data science in neurodegenerative disease: Its capabilities, limitations, and perspectives. *Current opinion in neurology*, 33(2): 249.
<https://doi.org/10.1097/WCO.0000000000000795>

11. Golriz Khatami S., **Mubeen S.**, Bharadhwaj S. V., Kodamullil A.T., Hofmann-Apitius M., and Domingo-Fernández D. (2020). Using predictive machine learning models for drug response simulation by calibrating patient-specific pathway signatures. *npj Systems Biology and Applications*, 7(1): 40.
<https://doi.org/10.1038/s41540-021-00199-1>
12. Bharadhwaj S. V., Mehdi C. T., Birkenbihl C., **Mubeen S.**, Lehmann J., Hofmann-Apitius M., Hoyt C. T., and Domingo-Fernández D. (2021). CLEP: A Hybrid Data- and Knowledge-Driven Framework for Generating Patient Representations. *Bioinformatics*, 37(19): 3311-3318.
<https://doi.org/10.1093/bioinformatics/btab340>
13. Golriz Khatami S., Russo M. F., Domingo-Fernández D., Zaliani A., **Mubeen S.**, Gadiya Y., ... Hofmann-Apitius M. (2021). Curating, collecting, and cataloguing global COVID-19 datasets for the aim of predicting personalized risk. *medRxiv*, 2021.11.14.21265797.
<https://doi.org/10.1101/2021.11.14.21265797>
14. Golriz Khatami S., Domingo-Fernández D., **Mubeen S.**, Hoyt C. T., Robinson C., Karki R., ... Hofmann-Apitius M. (2021). A Systems Biology Approach for Hypothesizing the Effect of Genetic Variants on Neuroimaging Features in Alzheimer's Disease. *Journal of Alzheimer's Disease*, 80(2): 831-840.
<https://doi.org/10.3233/JAD-201397>
15. Domingo-Fernández D., Gadiya Y., Patel A., **Mubeen S.**, Rivas-Barragan D., Diana C., ... Colluru V. (2021). Causal reasoning over knowledge graphs leveraging drug-perturbed and disease-specific transcriptomic signatures for drug discovery. *PLoS computational biology*, 18(2): e1009909.
<https://doi.org/10.1371/journal.pcbi.1009909>
16. Bharadhwaj, V. S., **Mubeen, S.**, Sargsyan, A., Jose, G. M., Geissler, S., ... Kodamullil A.T. (2022). Integrative analysis to identify shared mechanisms between schizophrenia and bipolar disorder and their comorbidities. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 122(1): 110688.
<https://doi.org/10.1016/j.pnpbp.2022.110688>

17. **Mubeen, S.**, Domingo-Fernández D., Díaz del Ser S., Solanki, D., Kodamullil A.T., Hofmann-Apitius M., Hopp M.T.,[†] and Imhof D.[†] (2022). Exploring the complex network of heme-triggered effects on the blood coagulation system. *Journal of Clinical Medicine*. 11(19), 597.
<https://doi.org/10.3390/jcm11195975>

Contents

1	Introduction	1
1.1	Biological data representations	2
1.1.1	Biological data	2
1.1.2	Molecular interactions	6
1.1.3	Interactomes	7
1.1.4	Context-specificity of interactions	8
1.2	Biological pathways	8
1.2.1	Pathway databases	9
1.2.2	Standard formats for interaction data	11
1.2.3	Enabling interoperability across formats	12
1.2.4	Pathway analysis	13
1.3	Network biology	14
1.3.1	Graph theory and definitions	14
1.3.2	Knowledge graphs	15
1.3.3	Algorithmic usage of knowledge graphs and networks	16
1.4	Organization and aims of this thesis	21

2	Interpreting the results of pathway enrichment analysis	24
2.1	On the influence of several factors on pathway enrichment analysis	24
2.2	The impact of pathway database choice on statistical enrichment analysis and predictive modeling.	27
2.3	DecoPath: a web application for decoding pathway enrichment analysis	30
3	Revealing context-specific expression patterns through integrated biological networks	33
3.1	Towards a global investigation of transcriptomic signatures through co-expression networks and pathway knowledge for the identification of disease mechanisms	33
3.2	Elucidating gene expression patterns across multiple biological contexts through a large-scale investigation of transcriptomic datasets	36
4	Network-based algorithms for biological applications	39
4.1	MultiPaths: a Python framework for analyzing multi-layer biological networks using diffusion algorithms	39
4.2	Drug2ways: reasoning over causal paths in biological networks for drug discovery	42
5	Conclusion and outlook	45
5.1	Future outlook	48
	References	50
A	Appendix	65

CHAPTER 1

Introduction

It has been long understood that living organisms are complex, composed of highly connected and interdependent systems working in concert to give rise to unique biological features and carry out discrete biological functions. However, system-wide studies tended to remain largely theoretical until recent decades. Now, state-of-the-art technologies are able to generate data in quantities that were once inconceivable by traditional approaches and which have finally made system-level investigations possible. With this vast volume of biological data, component parts (e.g., nucleic acids, proteins, and lipids) can be pieced together to elucidate how biological entities interact to carry out coordinated functions and how a system might fail and cause disease.

Many computational approaches have been developed for the study of biological data and among them, network-based approaches have become major ones. For instance, proteins are essential to nearly all cellular and molecular processes and only rarely do they act in isolation. By sequences of molecular events, such as the phosphorylation of a protein by another protein kinase, protein interactions form the scaffold for cellular responses, such as ensuring cell growth at a normal pace, unlike the abnormal rate in cancer [1]. By modelling these interactions, such as in protein protein interaction networks, we can improve our understanding of protein functions and cellular responses, elucidate disease etiology arising from aberrant functioning, and discover targets for disease.

This chapter introduces several concepts related to the types, measurements, modelling, storage and analyses of biological data. In particular, models which represent biological data within the context of their interactions

and associations are given special focus, as are pathway and network-level analyses of various types of *-omics* data in the biomedical domain. The publications which follow in later chapters detail the challenges in understanding complex biological systems, outline techniques for the interpretation of biological data and present methods for their investigation.

1.1 Biological data representations

1.1.1 Biological data

Advanced technologies have accelerated the rate at which biological data is produced. Next generation sequencing (NGS), microarrays, and mass spectrometry (MS) are among the major technologies used to characterize and quantify the complete (or partially complete) profiles of distinct classes of biological entities. With these developments, different stages in the transfer of genetic information, from DNA to RNA to protein, can now be holistically and rapidly measured along with other biological data modalities, such as metabolites [2]. The study of these profiles for a particular investigated molecular space is referred to as an *omics* study, providing a global survey of the state of a certain type of entity at any given point in time.

Omics studies are especially valuable as investigating the complete profiles of biological entities can help to elucidate normal cellular functioning and processes that lead to observable phenotypes. Furthermore, perturbations to a system can be evident in different ways at the molecular level. Investigating the alterations that are caused by such perturbations can help to decipher the etiology of disease. One can, for example, characterize the complete expression profile of a sample in various contexts and/or conditions and ask how expression varies from one context to another, whether there are differences between conditions, or how a perturbation, such as by a drug, can affect a system at large. Currently, several branches in this field investigate various biological data types, the most common of which are introduced below [3–6].

- **Genomics.** The field of genomics was the first of many *omics* disciplines to emerge. Focusing on the study of the complete genetic information of an organism (i.e., its genome), genomic experiments are used to

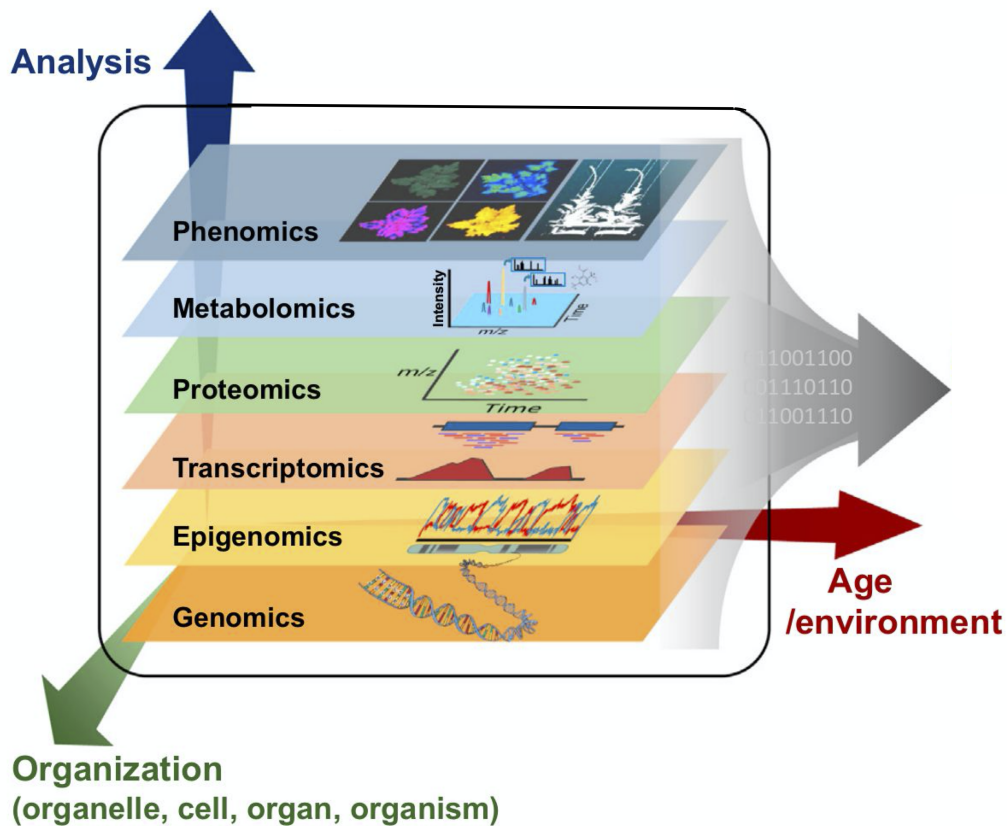


Figure 1: Omics disciplines. Omics studies at increasingly complex levels of biological organization. Image source: taken from [7].

identify associations between genetic variants and diseases as well as other phenotypes. The genome itself includes coding regions, genes which make up 1-2% of the entire genome, and non-coding regions, constituting the remaining 98-99% of DNA. Technologies associated with genomic data include DNA microarrays as well as NGS, such as whole genome sequencing (WGS), whole exome sequencing (WES), chromatin immunoprecipitation followed by sequencing (ChIP-seq) and single cell DNA sequencing (DNA-Seq). With these technologies, variations in the genome, such as single nucleotide variants (SNVs), insertions and deletions (indels), inversions, and copy number variations (CNVs) are routinely characterized. These data can be denoted by binary values, indicating that a gene is either wild type or mutated. While the vast majority of variants are benign, others may influence susceptibility to a disease or cause disease altogether, signifying the critical importance of their examination.

- **Epigenomics.** While genomics focuses on sequence data, epigenomics is the study of the epigenome, specifically, all chemical modifications to the genome that do not change the nucleotide sequence itself. This particular *omics* field is intended to investigate epigenetic modifications that play a key role in gene regulation (e.g., DNA methylation and histone modifications) as well as other processes. Notably, altered DNA methylation profiles have been noted in many diseases and can be used as disease biomarkers, for example in cancers, neuro-developmental disorders, metabolic disorders and autoimmune diseases [8]. Technologies for epigenomics include NGS (e.g., ChIP-seq) and array-based ones.
- **Transcriptomics.** Transcriptomics is concerned with the steady-state level of all mRNA transcripts of each gene. In this case, RNA sequences (or signals) are quantified and the abundance of RNA transcripts, or gene expression levels, are measured using technologies such as microarrays, RNAseq and single cell RNA-Seq (scRNAseq). The complete set of transcripts in a cell are collectively known as the transcriptome, including coding and non-coding RNAs, with coding RNA (mRNA) comprising 1-4% of the transcriptome. It is important to note that although RNA transcript or gene expression levels are often used as proxies to infer protein expression levels and gene activity, they may not accurately reflect either of these [9]. Although bulk RNAseq has been the predominant technology used to measure average global gene expression, bulk transcriptomic data can mask heterogeneity in cellular composition and cell types may also be sampled in varying proportions [10]. To offset these potential confounds and more accurately characterize the functional repertoire of individual cells, technological advancements within the past decade have resulted in the exponential scaling of scRNAseq experiments, now allowing for the parallel profiling of hundreds of thousands of individual cells in a single study [11] that can be stored in cell atlases, such as the Human Cell Atlas [12].
- **Proteomics.** Proteomics focuses on the study of the proteome, the complete set of proteins expressed in an organism, cell or tissue at a given point in time. Technologies, such as array-based and MS (although, primarily the latter) quantify protein abundance. The functions of many proteins are regulated by post-translational modifications (ex. phosphorylation, methylation, acetylation). These modifications affect various biological processes, such as the regulation of gene expression, signal transduction and DNA repair, and play key roles in various protein functions, including regulating enzyme activity and cell structure

maintenance [13]. Due to these modifications and additional factors, protein abundances can differ from the abundance of RNA transcripts, leading to some inaccuracies in inferring protein expression levels from gene expression levels. Compared to genomics and transcriptomics, proteomics tends to be a far more complex investigation owing to the large number of possible combinations of amino acids, polypeptide conformations and post-translational modifications of the resulting, functional protein. Finally, a single gene can encode for many proteins (e.g., due to alternative splicing), leading to a discrepancy between the total number of proteins (which tend to be much greater) than the total number of genes.

- **Metabolomics.** Metabolomics is concerned with the study of the metabolome, all metabolites which are the small molecule substrates and products of cellular processes in a biological system (e.g., carbohydrates, lipids, amino acids). Metabolites play a key role in cell functioning, including signal transduction and energy production. These molecules (e.g., ATP, acetyl-CoA) can regulate post-translational modifications (such as those described above) which affect protein activity, while the interaction of some metabolites with proteins can enable cellular responses through the initiation of signalling cascades [14]. When metabolite concentrations are outside of the normal range, these can be indicative of aberrant states. Analytical techniques associated with metabolomics include nuclear magnetic resonance (NMR) and MS-based technologies, which can be used to quantify all measurable metabolites, including unknown ones [15].
- **Other biological data types.** Another rapidly expanding field is microbiomics, the study of microorganism communities (such as those of bacteria, viruses, fungi and their genes) of a particular system. The human microbiota alone is estimated to contain 38 trillion bacteria, alluding to the complexity of the microbiome [16]. The impact of these microbes on human health are increasingly being discovered, with roles in cancer, disease susceptibility, infant health, anxiety, mood, cognition, and pain, amongst other indications [17, 18]. Common techniques associated with microbiomics include 16S ribosomal (rRNA) sequencing and shotgun metagenomics sequencing to extract DNA from microbial samples. Besides discrete biological entities, fields such as radiomics also exist for the examination of other biological data types, such as medical images from radiological modalities (e.g., computed tomography (CT), magnetic resonance (MRI), positron emission tomography (PET)), that

can be converted into high-dimensional data for quantitative feature extraction [19].

While *omics* studies have provided insights that have already been put into clinical practice, (e.g., the identification of disease biomarkers [20, 21] and the characterization of disease progression [22]), the disparate study of *omics* data modalities can fail to comprehensively capture the complexity of biological systems, which essentially can occur as interactions of discrete biological entities across multiple *omics* layers [23]. This has led to the emergence of multi-*omics* experiments, where multiple *omics* technologies are combined to generate integrated *omics* data [3, 24, 25]. Several resources and platforms which integrate diverse *omics* data types have become available, such as ColPortal for methylation, transcriptomic, microbiomic and clinical data, among other types, for colon cancer patients [26, 27].

1.1.2 Molecular interactions

Individual biological entities, such as those described in the preceding section, rarely act in isolation to carry out biological functions. Instead, interactions, referring to the physical or functional association of two biological entities which cause or result in some biological effect, are the main drivers of cellular and molecular functions. These interactions can occur between the same types of entities (e.g., protein-protein interactions), across modalities (e.g., protein-metabolite interactions) as well as across biological constructs (e.g., gene-phenotype interactions) and can be divided into various classes, some of which are described below [28–30].

- **Protein-protein interactions** (PPIs) are stable or transient physical interactions between proteins which can result in the formation of a protein complex or a specific, temporary response, respectively. PPIs can be experimentally obtained via methods such as yeast-2-hybrid assays, MS-based approaches, or co-immunoprecipitation, while computational approaches for obtaining PPIs include prediction methods (e.g., homology-based interaction inference) or literature text-mining. PPI interaction resources include STRING [31], BioGrid [32] and APID [33].
- **Regulatory interactions** refer to the binding of proteins (i.e., transcription factors (TFs)) to certain regions of DNA which enhance or

repress the expression of one or more genes and cause a change in their activity. Various experimental methods can be used to characterize TF-DNA binding, such as ChIP-seq, while databases which house these interactions include ReMap [34] and TRUUST [35].

- **Metabolic reactions** are biochemical interactions that occur between metabolites and enzymes. These reactions result in the conversion of metabolites from one form to another, with each step in the conversion process mediated by specific enzymes, which catalyze these reactions. Databases of metabolic reactions include KEGG [36], MetaCyc [37], ENZYME [38] and BRENDA [39].
- **Other types of interactions.** Other types of molecular interactions include signalling reactions, where post-translational modifications of one protein by another initiate a biological signal or transmit a signalling cascade, and interactions between metabolites and proteins, such as a drug and its protein target, where the metabolite can cause some alteration to the activity of the protein. However, these protein-metabolite interactions can be difficult to characterize due to their transient nature and weak affinities. MS- and NMR- based approaches have nonetheless been used to systematically map protein-metabolite interactions [40, 41].

Various disease states can occur in the event of disruptions to normal interaction behaviours. For example, certain mutations in transcription factors (e.g., gene amplification, gene deletions and point mutations) can modify the regulatory circuits of a cell and are known to contribute to several diseases (e.g., cancers, autoimmune disorders, cardiovascular diseases and diabetes) [42–44]. An example lies with the oncogenic transcription factor, TAL1, whose binding with GATA3 and RUNX1 form a positive autoregulatory loop, driving an altered circuitry that likely contributes to the oncogenic human T cell acute lymphoblastic leukemia program [45].

1.1.3 Interactomes

Sets of molecular interactions can be arranged into what is known as an interactome, a network of molecular interactions between biological entities. Generally, the interactome is most often used to refer to sets of PPIs, such as the human interactome for all PPIs in humans cells. Nonetheless, interactomes

of entities from other *omics* layers have also been established, such as the human protein-DNA interactome (i.e., gene regulatory network) [46], RNA interactome [47], protein-RNA network [48], as well as the gene-interaction network of indirect, functional relationships between genes. By generating these interactomes, it becomes possible to model complete networks of interactions between individual biological entities to facilitate an understanding of their collective roles in normal and aberrant cellular functions.

1.1.4 Context-specificity of interactions

While the elucidation of molecular interactions and their representation within interactome networks is a significant step in modelling the complexity and interplay of entities within a functional, biological system, an important caveat is that interactions, and by extension, interactomes, tend to be void of biological context. For instance, in a study by Stacey *et al.* [49], the authors found no evidence for the occurrence of anywhere between 19 to 55% of interactions reported in several literature-curated PPI databases. Thus, a molecular interaction may only be a snapshot of an event occurring in a particular cell and/or condition at a given point in time.

Although the cells of an organism contain the same DNA sequence, different cells can exhibit distinct behaviours and characteristics (e.g., a neuron vs. a white blood cell). These differences result from the specific binding of TFs to particular positions on the DNA sequence, resulting in variable levels of expression of particular genes and the diversity observed across different cell types. Consequently, this can mean that the conditions in which an experimental molecular interaction occurred may not be reflective of the actual interaction occurring within a given context (e.g., cell type or tissue).

1.2 Biological pathways

While the interactomes described in the preceding section comprised sets of all interacting biological entities, a more knowledge-based approach to representing biological data is by collating sets of interactions into biological pathways. These pathways are essentially series of interactions between physical biological entities (or biological constructs) which result in a particular

event, such as a cellular change or the formation of a product [50]. Collectively, pathways carry out some biological process (see **Figure 2**). For example, a biological pathway can represent how a signal is transmitted from an external to an internal environment, a particular carbohydrate is metabolized, DNA is repaired, or how a pathogen affects a host cell. Types of pathways which can be modelled include metabolic, signal transduction, gene regulation and disease pathways.

Not surprisingly, representing biological data in this simplified abstraction does imply some loss of information, such as spatio-temporal features. Furthermore, pathway representations can include interactions characterized in disease states *in vitro* (e.g., HeLa cancer cell line), which although may be appropriate for the study of disease pathways, may not be generalizable to normal states. For example, interactions characterized in immortalized or tumorigenic fibroblasts in cell signalling pathways can be inherently biased and ill-suited to represent normal cellular functioning. Despite these shortcomings, pathways are advantaged by their capacity to formalize and abstract well-established relationships with literature validation, and by their ability to facilitate a visual and more easily interpretable understanding of complex processes. They are additionally advantaged by their capability to represent biological entities and constructs across multiple scales (e.g., from genotype to phenotype), and encode rich relationship descriptions (e.g., activation, inhibition, binding/association, phosphorylation) [51].

Deconstructing the mechanisms by which a particular biological function is executed, not at the level of individual entities, but rather, considering their interplay as a whole, can serve several advantages. This includes understanding how individual entities assemble into modules to carry out discrete functions, as well as disease etiology. For example, this component mapping can help to identify pathways that are dysregulated by a disease as well as affected routes along specific pathways in order to reveal potential therapeutic targets.

1.2.1 Pathway databases

Given the popularity of pathway maps for the representation of biological data, a particular class of databases have emerged to formalize, collect and store biological pathways. Several academic and commercial enterprises have produced pathway databases numbering in the hundreds, reflecting the

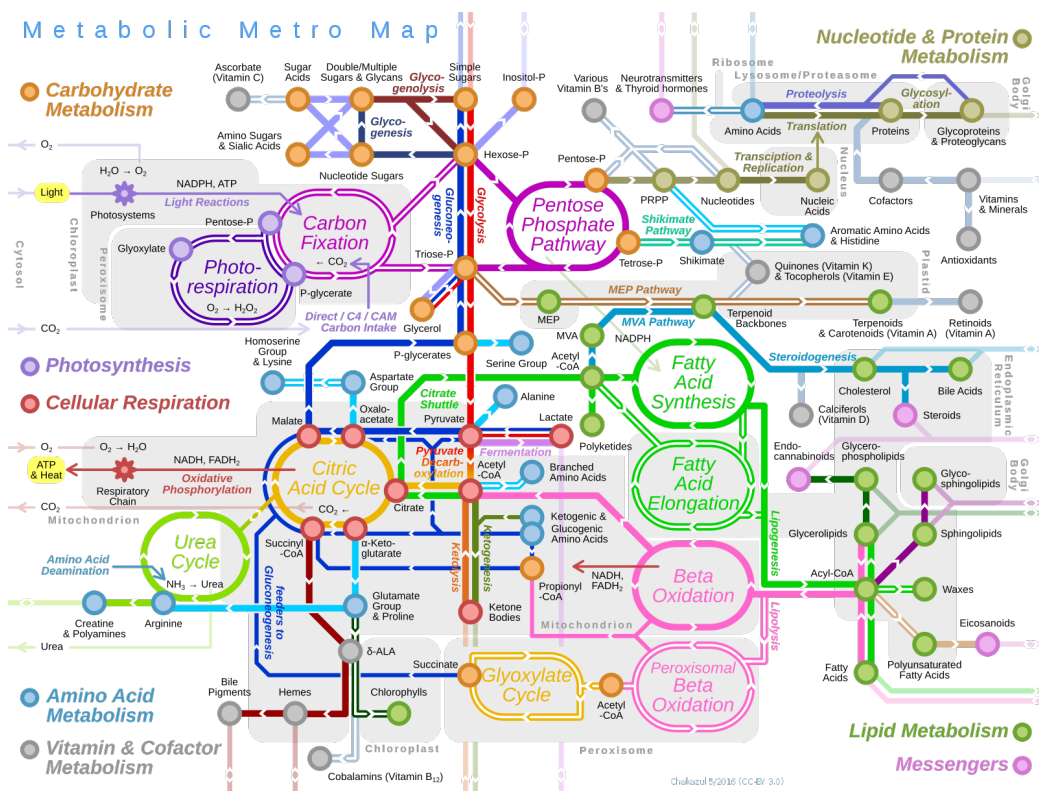


Figure 2: Biological pathway map. Image source: taken from [52].

diversity of biological processes occurring in living organisms [53].

Many of these databases can be distinguished by the domains that they cover, with some pertaining to a particular species (e.g., EcoCyc [54], Plant Reactome [55]), or disease (e.g., NeuroMMSig [56], AlzPathway [57], Atlas of Cancer Signalling Network (ACSN) [58]), while others differ by their content, such as those focused on signalling (e.g., SIGNOR [59], NetPath [60]) or metabolic pathways (e.g., MetaCyc [37], BRENDA [39]). Additionally, some databases can be highly specific (e.g., YTRP [61] for transcriptional regulatory pathways in *Saccharomyces cerevisiae*), while others are more comprehensive, covering hundreds or thousands of pathways across multiple species, such as KEGG [36], Reactome [62], WikiPathways [63], and PathBank [64].

Besides differences in their content, pathway databases can also be distinguished by several other factors. For example, pathways across databases can be described in varying levels of detail such that some biological entities and/or interactions may be included within a pathway in a particular database, but excluded in the same pathway in another database [65, 66]. Additionally,

the average number of pathways in a given database can range from several hundred (e.g., KEGG), to over a hundred thousand (e.g., PathBank), hinting at variable definitions of pathway boundaries. The interconnected nature of biological pathways (as alluded to in **Figure 2**) also implies arbitrary definitions of pathway boundaries, which can be variously selected by domain experts for disparate databases. For instance, one database may define a particular biological process as a single pathway, while another database may define several parts of that same biological process as individual functional modules and separate pathways. One possible solution to standardize pathway definitions lies with a technique proposed by Belinky *et al.*, [67] that uses hierarchical clustering and nearest neighbour graph representation to group similar pathways. Although the approach they use has been intended for merging pathways, it can also be used as an objective standard to define pathway boundaries, where boundaries are drawn to minimize redundancy across associated gene sets. In summary, even well-established, canonical pathways can differ across databases due to the aforementioned sources of variability, implying a degree of subjectivity in abstracting pathway knowledge.

1.2.2 Standard formats for interaction data

Pathways are essentially computational data models which describe interactions between heterogeneous biological entities and/or constructs. In order to transform interaction data into pathways, several languages have been made available and are variously used by disparate pathway databases. Below, we provide a brief overview of various formats which have been adopted by the scientific community.

Among the most commonly used interaction formats are BioPAX and SBML. Biological Exchange, or BioPAX, is a standard RDF/OWL- based language that can be used for the representation of various molecular and genetic interactions, pathways, and gene regulatory networks [68]. BioPAX is established as a major pathway exchange format for the integration, visualization and analysis of pathway data. Databases which offer BioPAX export include Reactome, MetaCyc and WikiPathways. The Systems Biology Markup Language (SBML) is a standard XML-based language that is used to represent computational models of system biology [69]. Much like BioPAX, SBML is also intended as an exchange format and to describe various biological processes, such as metabolic and signalling pathways, with a particular focus on representing biochemical network models. Reactome, BRENDA,

MetaCyc and PANTHER are among the databases which provide SBML export.

Apart from the languages mentioned thus far, additional interaction formats that have been widely used include BEL, SBGN, PSI-MI, RDF and SIF. Similar to the above mentioned languages, Biological Expression Language (BEL) formalizes biological relationships in a computable form, and also includes rich, contextual-descriptions of causal and correlative relationships across biological scales [70]. Systems Biology Graphical Notation (SBGN) has also been developed as an unambiguous standard for the representation, storage, visualization and exchange of biological processes [71], while the Proteomics Standards Initiative Molecular Interactions (PSI-MI) [72] exchange format is especially popular among databases of protein-protein interactions, such as IntAct [73], BioGRID [74] and MINT [75]. Finally, the triple structure (i.e., subject, predicate and object) of the Resource Description Framework (RDF) is also used as a standard model to represent biological relationships [76], while the Simple Interaction Format (SIF) can be used to generate networks from lists of molecular interactions.

1.2.3 Enabling interoperability across formats

Although collectively, pathway databases cover a broad scope of information in varying levels of detail, due to a diversity of formats and the lack of interoperability between them, they tend to be fairly disconnected and only independently accessible to researchers. In addition, standardized nomenclature for pathways are lacking, further hindering pathway interoperability. Although some efforts have been made to standardize pathway nomenclature, as with the Pathway Ontology for the annotation of genes to pathway terms for multiple species [77], these have not yet been widely adopted. Despite these challenges, several software converters have been developed for the translation of one language to another (e.g., PathMe (**Figure 3**) [66], PAX2GRAPHML [78], PaxTools [79]). Furthermore, meta-data databases, such as PathwayCommons [80] and ConsensusPathDb [81], have also been established to consolidate pathway knowledge and interaction data across several primary resources and represent data in a standardized format. These integrative approaches can serve to provide a far more holistic view of the knowledge surrounding a particular process and more accurately represent literature findings as opposed to the knowledge any singular resource may accumulate.

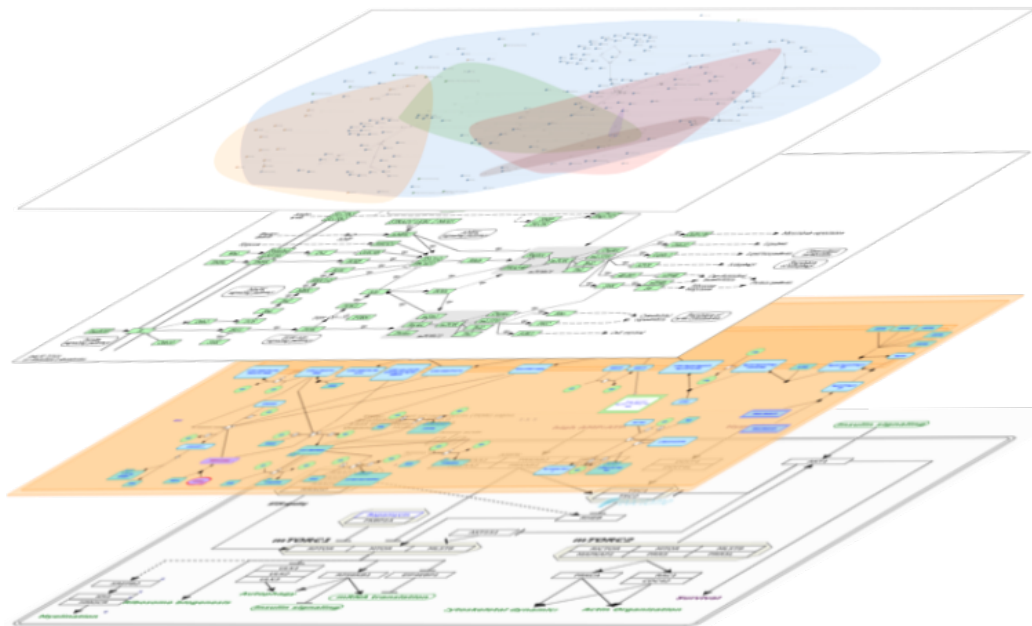


Figure 3: Pathway overlay. Illustrations of the mTOR signaling pathway in multiple pathway databases (i.e., KEGG, Reactome and WikiPathways) depicted in the three lowest layers, and their merged representation depicted in the topmost layer.

1.2.4 Pathway analysis

A prototypical method for the interpretation of high dimensional biological data has become pathway enrichment analysis, a term which encapsulates a group of analytical methods that investigate whether a pathway is enriched, or over-represented, within a list of genes [82]. Typically, these genes are derived from a high throughput experimental dataset associated with a given phenotype. These analyses are primarily intended to garner mechanistic insights on large volumes of biological data by using pathway representations that can summarize high dimensional information (e.g., thousands of genes) to a handful of biological processes. A question that is commonly addressed by a pathway analysis is whether specific sets of genes may be associated with a given phenotype or biological process.

Despite advances in available pathway enrichment methods, the vast majority discard topological pathway information. Instead, a pathway is simplified to a set of genes without any interaction information (i.e., a gene set), and the gene set is tested against differentially expressed genes from an experimental dataset that investigates a particular phenotype (e.g., breast

cancer versus normal) [83]. More specifically, a common pipeline for several enrichment methods is to first obtain a ranking of genes according to their degree of differential expression. Each gene set within a collection or from a pathway database is then tested to determine whether genes in the gene set are over-represented at the top and/or bottom of the ranked list of genes. If the genes of the gene set are clustered at either the top and/or bottom of the ranked list more so than expected by chance, the gene set may be statistically significant for the particular phenotype under study. If, however, the genes within the gene set are equally distributed throughout the ranked list, this suggests that the gene set is unlikely to be interesting or relevant in some statistically significant way to the investigated phenotype [84, 85]. The results procured from such an analysis can thus shed light on biological processes that may be affected by a particular condition which may otherwise go unnoticed if an experimental dataset were to be examined solely at the gene level rather than the pathway level.

1.3 Network biology

The networks described thus far (e.g., the interactome and biological pathways) all fall within the framework of network biology, the abstraction and mathematical depiction of relationships between biological entities and/or concepts [86]. The field concerned with the modelling of pairwise relationships between objects such as these is formally referred to as graph theory. In the sections that follow, we introduce key definitions and concepts within graph theory, as well as network-based methods and their applications in computational network biology.

1.3.1 Graph theory and definitions

A graph can be defined as $G = (V, E)$, where V is a set of vertices that represent nodes and E is a set of edges that represent connections between nodes. Standard data structures to represent graphs are through collections of adjacency lists or through adjacency matrices. Adjacency list-based representations are particularly suited for sparse graphs, though an adjacency matrix representation may be preferred in cases when a graph is dense, or for efficient lookup of edge connections between specific vertices. There exist

several classes of graphs which are typically defined by their edge types, as described in Table 1 [87–89].

Of the types of molecular interactions described in subsection 1.1.2, PPIs are typically modelled in undirected graphs, representing symmetric binding relationships, while metabolic reactions, signalling reactions, and regulatory interactions are frequently modelled as causal edges in directed graphs. Edges in an undirected weighted graph can be used to model the strength of correlated expression of genes in one type of biological network, termed co-expression networks [90]. Applications of bipartite graphs can include the representation of enzyme-reaction links, gene-disease links and drug-target links [91]. Finally, biological applications of hypergraphs can include the modelling of metabolic reactions, where many substrates are converted into many products.

1.3.2 Knowledge graphs

While biological pathways represent the series of interactions that occur between biological entities and/or constructs which carry out some biological process, yet another abstraction of biological entities and constructs, either sequential and/or descriptive, is in what is known as a knowledge graph (KG). Formally, a KG is a directed, labelled graph in which relations have labels with logical, well-defined meaning and which graphically structure the knowledge within a particular domain [92, 93]. In the biological domain, KGs can be further characterized as modelling heterogeneous relationships (e.g., activation, inhibition, methylation) across a wide range of biological scales, including physical entities (e.g., genes, proteins, and metabolites) and higher order concepts (e.g., biological processes, phenotypes, and diseases). KGs can be constructed from several resources, such as databases (e.g., KEGG, DrugBank [94], STRING), ontologies (e.g., GO), through manual curation of the literature or through text mining [95, 96].

Applications of KGs typically involve reasoning over a KG to study a particular hypothesis [97]. For instance, several studies have leveraged KGs for the prediction of drug-drug [98–103], drug-target [104] and drug-side effect interactions [105, 106], as well as to infer drug-disease associations [107]. Additional applications have included the discovery of antibiotic resistant *E. coli* genes through a KG of antibiotic resistance in *E. coli* [108], patient diagnoses and treatment recommendations for clinical support, information

Graph	Definition
Undirected	A graph G is undirected if a pair of vertices $(u, v) \in E$ are neighbours with no assigned direction (Figure 4a).
Directed	A graph G is directed if vertices in edges are ordered. A directed edge $E = (u, v)$ is considered to have direction from u to v (Figure 4b).
Weighted	A graph G is weighted if each edge has an associated weight, generally given by a weight function $w : E \rightarrow \mathbb{R}$.
Bipartite	A graph G is a bipartite graph if V can be partitioned into 2 sets V_1 and V_2 such that for each edge (u, v) in E , $u \in V_1$ and $v \in V_2$ or $v \in V_1$ and $u \in V_2$.
Complete	A complete graph is an undirected graph where every pair of vertices is connected by a unique edge.
Directed acyclic	A directed acyclic graph is a directed graph which does not contain cycles.
Hypergraph	A graph H is a hypergraph if it contains a set of vertices V and a set of hyperedges E , where a hyperedge can join an arbitrary number of vertices, unlike a simple edge which joins exactly two.
Tree	A tree is a type of undirected graph where each pair of vertices is connected by exactly one simple path.

Table 1: Survey of various graph types.

retrieval from medical reports and the prediction of medication prescriptions via a KG constructed from electronic medical records [109].

1.3.3 Algorithmic usage of knowledge graphs and networks

Various network-based methods have broadly been used for biological applications. In what follows, we describe some major categories of approaches and algorithms which have been applied in the biomedical domain.

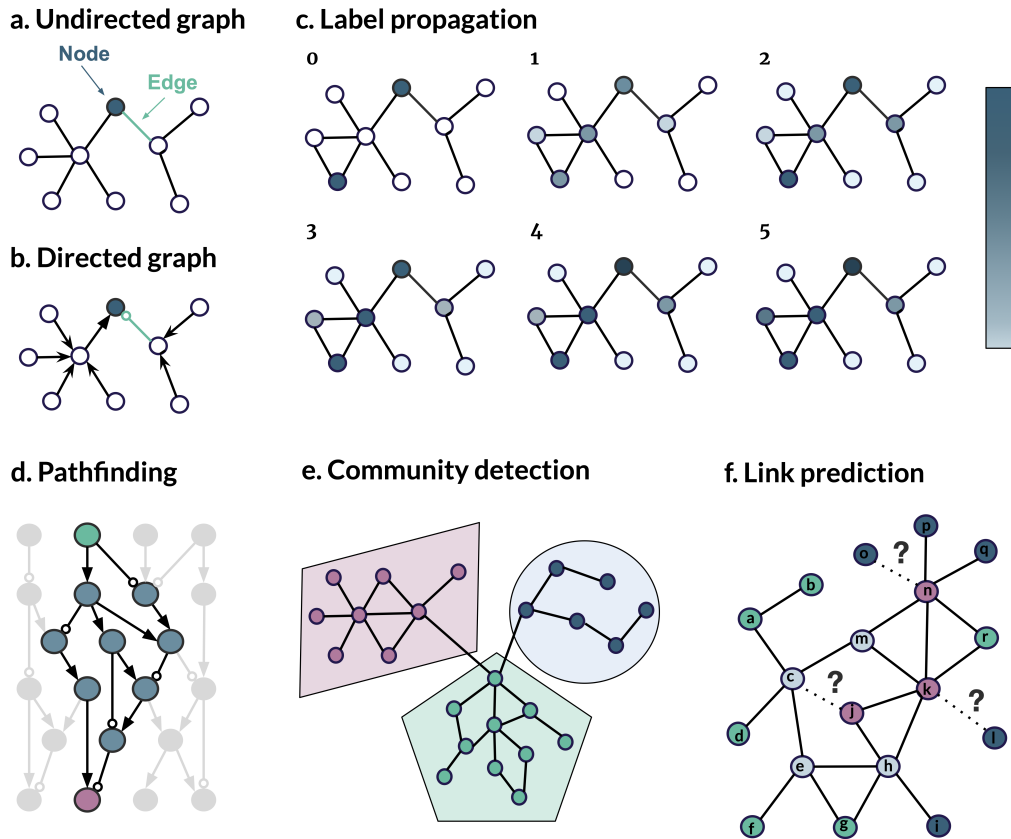


Figure 4: Graph representations and network-based methods. Example design choices of graph representations include (a) undirected and (b) directed graphs. (c) An undirected graph illustrating the propagation of scores by a label propagation algorithm, where nodes with high initial scores are coloured. Over a series of iterations (0-5), the scores are propagated to neighbouring nodes before reaching convergence. (d) A directed graph depicting a pathfinding task from a source node (labelled green) to a target node (labelled purple) through all paths between the two nodes. (e) An undirected graph highlighting a likely topological community structure, in which nodes within a community have a greater likelihood of being highly connected, while their connections to other groups are more sparse. (f) An undirected graph in which specific links (dashed lines) denote additional possible edges which a link prediction algorithm is tasked with predicting. (c) and (d) have been adapted from [110] and [111], respectively.

- Label propagation.** The principle underlying label propagation is that nodes in close proximity within a network tend to have community structure and can share common attributes [110]. On account of this principle, given a partially labelled graph, a label propagation algorithm

assigns unlabelled nodes within a network with labels, assuming that the propagation will subsume nodes within a community. Through several iterations of the algorithm, known labels are propagated or diffused to neighbouring nodes within the network (**see Figure 4c**) [112], which, if unlabelled, are inferred. Biological applications for label propagation include protein function prediction (e.g., GeneMANIA [113]), disease module detection (e.g., HotNet2 [114]) and patient stratification (e.g., Similarity Network Fusion (SNF) [115]).

- **Pathfinding.** The task of pathfinding is concerned with exploring paths between nodes in a graph. Beginning at some start node, algorithms for pathfinding traverse along relationships through adjacent nodes for general exploration, or until an explicit destination is reached [116] (**see Figure 4d**). This can be achieved by brute force approaches, such as breadth-first or depth-first search, through heuristics or dynamic programming. By incorporating information on node distances and traversal direction, pathfinding algorithms often search for the most cost-effective path, such as the optimal shortest path. Within a biological context, shortest path approaches have been used for various applications in network medicine. For instance, studies have used shortest path approaches for drug repurposing (i.e., associating an existing drug to a disease for which the drug was not previously known to be therapeutic towards) [117–119] under the premise that drugs that are structurally similar regulate proteins in close proximity within a PPI network [120]. By contrast, the problem of finding longer paths between nodes within a network quickly becomes intractable due to an exponential increase in the number of possible paths as the number of nodes increase. Nonetheless, longer paths can also be considered by constraining them via meta-paths or by leveraging experimental *omics* data, as in [111].
- **Community detection.** Community detection aims at identifying topological community structure within networks [86]. These communities represent groups of nodes which have a greater likelihood of being connected to each other than nodes which are in other groups such that nodes within a group are densely connected, while their connections to other groups are sparse [121, 122] (**Figure 4e**). The study of communities is especially relevant in a biological context given that, i) biological networks are typically far too large to examine as a whole and, ii) a common feature of network communities is that they tend to correlate with specific biological functions [123, 124]. In one applica-

tion, using unsupervised Markov clustering [125] on an experimentally derived proteome interaction network, Huttlin *et al.* [126] identified 1,300 protein communities corresponding to diverse cellular functions and 442 communities associated with over 2,000 disease annotations, shedding light on potential candidate disease genes within the network.

- **Disease module detection.** Disease module identification represents a major application within network medicine [127]. In one study, the authors found that disease-relevant genes for 226 of nearly 300 investigated diseases were significantly more likely to appear in communities or disease modules [128]. Specifically, the examination of modules within a network have been used to identify potential, novel disease-relevant genes as those which are neighbours of known disease genes in a particular disease module. However, topological community detection algorithms are unable to directly define disease modules and instead require distinct computational approaches, such as incorporating *omics* information or first identifying some disease-related genes through empirical experimentation (e.g., DIAMOnD approach [129]) [86].
- **Link prediction.** A link prediction task is concerned with the prediction of new links, or the inference of missing ones, between pairs of nodes within a network based on existing links and node attributes [130] (**Figure 4f**). Algorithms for link prediction can be divided into three broad categories, specifically those which are similarity-based, machine learning-based, or probabilistic and statistical models. Of the three, similarity-based algorithms tends to be the most commonly used class of link prediction algorithms in network biology, assuming that links between nodes which are similar or close to each other in a network have a greater likelihood of occurring [86]. Slightly deviating from the principle underlying this assumption, Kovács and colleagues [131] employ a similarity-based method that uses both local and global topological information to establish whether a link may exist between two unconnected protein nodes. Here, the authors assert that whether two proteins interact is determined not by whether they are similar enough to each other, but rather, whether one of them is similar enough to the other's established interaction partners. Link prediction approaches have also been applied for drug discovery, including predicting novel or missing links between drugs and their targets, specifically proteins, diseases and other drugs [132].

Machine learning in network biology

Machine learning (ML) is a branch of computer science and artificial intelligence (AI) concerned with learning patterns within datasets through a combination of mathematical rules and statistical assumptions [133]. A common task in ML can be briefly characterized as follows: i) a dataset of features across samples is processed, ii) based on a prediction task on the dataset, an ML approach is selected, iii) the model is trained on the known input data such that, iv) the trained model is primed to make predictions on new data. While potential applications of ML abound in network biology, manual labour and domain expertise are required to extract informative features from networks as inputs for ML models. In response, the field of Network Representation Learning (NRL) emerged, concerned with the automatic representation of graph structures in a Euclidean space that circumvented the need for manual feature engineering [134]. Using various NRL approaches, graphs could be taken as input and different modes of information within the graph, such as structural and biological information, could be preserved in a latent space [135].

The intended output of these approaches depends on the research question being asked, and can be vector representations at the level of nodes, edges or entire graphs. An example at the node-level entails learning how each node in a network can be mapped to a low-dimensional space and representing the nodes as vectors of d numbers (i.e., embeddings), where similar nodes in a network are close within the embedding space (e.g., similar nodes in the graph are embedded closer together) [86]. These node embeddings can subsequently be used for downstream ML/AI tasks, such as node classification to predict novel functions of protein nodes in a network. Analogously, edge embeddings generate low-dimensional vectors of edges within a network and can be used as feature inputs for tasks including the prediction of novel interactions in biological networks (**Figure 4f**). Finally, entire-graph embeddings can also be learnt for applications such as drug discovery by embedding graphs that represent entire molecules, and classifying these molecules to identify potential disease candidates [136]. An example of a graph embedding at the level of nodes and edges is illustrated in (**Figure 5**).

Methods which automatically learn to project graph structure into low dimensional embeddings can include those suitable for homogeneous networks, such as matrix factorization-based (e.g., Laplacian eigenmaps [137]), random walk-based (e.g., node2vec [138]) or deep learning-based (e.g., Graph Convolutional Networks [139]), and those suitable for heterogeneous networks (e.g.,

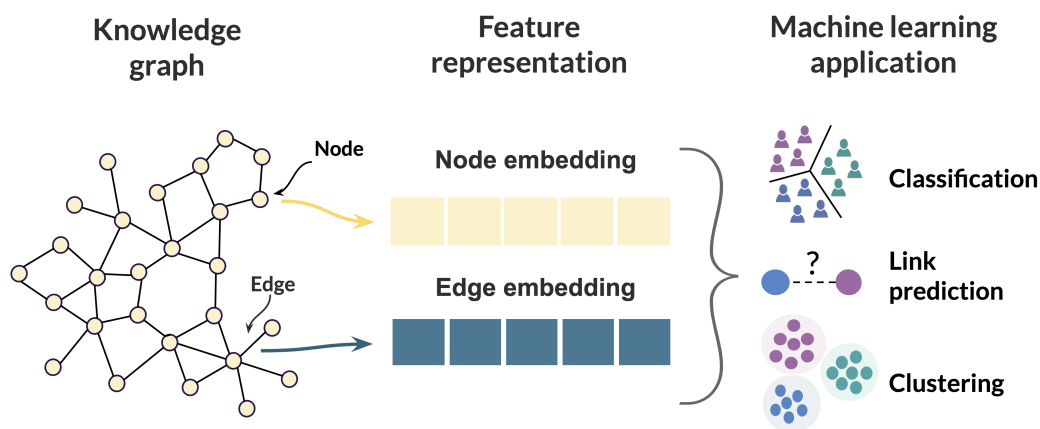


Figure 5: Node and edge embeddings. Given a KG as an input, learnt embeddings of nodes and edges within the KG can be used as feature inputs for various downstream ML/AI tasks, such as node classification, link prediction and node clustering.

KGs), such as semantic matching models (e.g., HolE [140]), translational distance models (e.g., TransE [141]) and meta-path based methods (e.g., metapath2vec [142]) [96, 143]. Biological applications of network embeddings have variously included the prediction of protein functions [144], drug repurposing candidates [145], gene-disease associations [146] and drug-targets [147], as well as the identification of pathways that mediate drug response [148] and the identification of polypharmacology side effects [106]. Similarly, language models in natural language processing (NLP) can be used to project words into embeddings [149]. For instance, in Balabin *et al.* (2022), the authors demonstrated how KG and text embeddings (i.e., word sequences transformed into vectors) can be combined for more robust predictions on multiple multi-class classification tasks in biological applications, such as, predicting the disease context within which a particular relation occurs [150].

1.4 Organization and aims of this thesis

The abundance of biological data that has so far been generated is as valuable to gaining biological insights as are the techniques used to interpret, contextualize and recognize meaningful patterns within it. A major technique to do so lies in modelling relationships between measured entities and biological concepts using networks, as in the field of network biology. In this work, we focus on network representations for the interpretation and contextualization of

high dimensional biological data, as well as demonstrate how domain-specific networks and knowledge graphs can be operationalized for algorithmic utility for various biomedical applications. Specifically, the work in this dissertation serves to address the three major objectives outlined below, each of which is the subject of the following chapters of this thesis.

- i. Shed light on existing challenges in the *interpretation* of high dimensional biomedical data and provide solutions to mitigate these challenges (Chapter 2).
- ii. *Contextualize* transcriptional patterns to reveal complex regulatory circuits that underlie discrete biological functions (Chapter 3).
- iii. Develop network-based algorithms for various biological applications, including prioritizing drugs for a given indication and disease module identification (Chapter 4).

In Chapter 2, we first present a comprehensive review of the current literature on pathway analysis methods that leverage biological pathway knowledge (as per Section 2.1). Specifically, we introduce and detail the factors that impact the results of pathway analysis and the interpretation of these results, and summarize the findings of major comparative studies that have benchmarked these factors. Of these major factors, in Section 2.2, we particularly focus on the choice of pathway database and design a series of experiments to evaluate the impact of this critical factor on the results of pathway enrichment analysis. In this work, we observed that different databases can indeed yield disparate results on enrichment analysis and even when the same pathway is represented in different databases, alternative pathway definitions can result in its significant enrichment or lack thereof. Finally, in concluding this chapter, we introduce DecoPath (as per Section 2.3), a web application which demonstrates how the integration of many pathway resources can consolidate knowledge surrounding known interactions and be an asset in a pathway analysis. DecoPath serves to facilitate the interpretation of disparate results, and illuminate findings from consolidated results for pathway analysis.

Then, in Chapter 3 we present two publications which introduce a specific class of biological networks in which nodes (genes) are connected depending on the strength of their correlations. This chapter emphasize the importance of considering specific contexts into account for the interpretation of biological data, namely disease context (as per Section 3.1) and additional contexts, including, cell types, tissues, and cell lines (as per Section 3.2).

Finally, in Chapter 4, we conclude with two publications that adopt methods from graph theory for biological applications. Specifically, these include MultiPaths (as per Section 4.1) and drug2ways (as per Section 4.2). While MultiPaths focuses on label propagation algorithms for multi-modal biological networks, drug2ways is an advanced algorithm that has been developed for pathfinding-based drug discovery and repurposing in knowledge graphs and complex networks.

These chapters are subsequently followed by a discussion of the themes presented, challenges faced herein, and possible future directions, serving as a general conclusion of this thesis.

CHAPTER 2

Interpreting the results of pathway enrichment analysis

One of the primary avenues researchers have for the interpretation of high dimensional biological data is pathway enrichment analysis to investigate the involvement of particular sets of genes in a given phenotype. In this chapter, we introduce a series of publications which reveal common challenges associated with interpreting the results of these analyses and establish techniques to mitigate these challenges.

2.1 On the influence of several factors on pathway enrichment analysis

This section presents the following publication (see **Appendix A.1**):

Sarah Mubeen, Alpha Tom Kodamullil, Martin Hofmann Apitius and Daniel Domingo Fernández (2022). On the influence of several factors on pathway enrichment analysis. *Briefings in Bioinformatics*, 23:3.

Summary

Among the techniques available for the interpretation of biological data, pathway enrichment analysis has emerged as one of the more prominent. The popularity of this type of analysis can be attributed to a more natural approach in which biological data is interpreted within a system-level context, guided by literature- and/or expert- based knowledge. Given its widespread popularity, several hundreds of enrichment methods and pathway databases have been developed, yet paradoxically, gold standards have noticeably been lacking.

Nonetheless, the lack of universal, gold standards is not surprising given the variability conferred by the variety of possible configurations, modular aspects and interchangeable factors possible when conducting a pathway analysis. For instance, an experimental dataset can possess different characteristics (e.g., varying number of samples or varying degrees of differential expression among entities from different experimental groups), choosing one pathway database or gene set collection over another can result in differing pathway definitions and gene set sizes, and a wide range of enrichment methods (e.g., topology versus non-topology -based) and configurations (e.g., different gene (i.e., local) and gene set-level (i.e., global) statistics) are possible. Consequently, these methods and databases have been investigated in a number of benchmark studies alongside of various other factors to study the impact they produce on the results of enrichment analysis.

The publication, *On the influence of several factors on pathway enrichment analysis* provides a comprehensive review of the literature on key factors of pathway analysis and summarizes the results of studies which have evaluated the influence of these factors. Solutions to mitigate the effect of these factors and identify possible future benchmarks are also made.

The study finds that in many instances, the results of enrichment analysis can be ascribed to various factors beyond the intended goal of investigation. We have reviewed numerous studies which have demonstrated how alternative experimental designs of such an analysis can lead to different results. In doing so, we have provided a comprehensive overview of which aspects a researcher should be cognizant of in conducting an enrichment analysis and how these can impact results. We have especially focused on a dozen studies, representing all major comparative studies performed to date which have collectively benchmarked nearly fifty enrichment methods and/or their

variants. Summarizing the findings of these studies, we have found severe inconsistencies across the performance of methods. In some extreme cases, different methods can go so far as to yield either all gene sets or no gene sets as significantly enriched on the same dataset and database, highlighting the importance of the careful consideration of these factors. Nonetheless, despite the inconsistencies noted, we were able to observe trends across studies with respect to the performance of methods on several metrics, such as specificity, sensitivity and prioritization. This revealed some methods do rank higher than others on a particular metric, although none was found to outperform all others across all metrics.

We have also critically reviewed several other major factors that can influence pathway enrichment analysis, including variations in modular aspects of a typical enrichment analysis, gene set size, and gene set/pathway database choice, once again finding these to be highly consequential to the overall results. The choice of reference pathway database is especially significant, given that the results of an enrichment analysis are determined by the pathways or gene sets included in the collection. Nonetheless, the choice itself can at times be arbitrary, selected due to popularity, ease of usage and prior experience. However, the influence of this factor can be mitigated by integrating multiple resources into the analysis instead of relying upon a single one for more comprehensive and less variable results, as we demonstrate in the section that follows. Finally, we have provided possible solutions to mitigate the outlined factors, proposed possible future benchmarks, and made recommendations for researchers to make well-informed decisions when conducting a pathway analysis.

Authors' contributions

Sarah Mubeen and Daniel Domingo Fernández wrote the manuscript. Alpha Tom Kodamullil and Martin Hofmann Apitius reviewed the manuscript.

2.2 The impact of pathway database choice on statistical enrichment analysis and predictive modeling.

This section presents the following publication (see **Appendix A.2**):

Sarah Mubeen, Charles Tapley Hoyt, André Gemünd, Martin Hofmann Apitius, Holger Fröhlich, and Daniel Domingo Fernández. (2019). The Impact of Pathway Database Choice on Statistical Enrichment Analysis and Predictive Modeling. *Frontiers in Genetics*, 10:1203.

Summary

As outlined in the previous section (2.1), there are several factors that influence the results of pathway enrichment analysis, including the enrichment method and database choice. While over a dozen benchmark studies have been conducted to explore the influence of enrichment methods, the choice of pathway database has received far less attention. In the publication, *The Impact of Pathway Database Choice on Statistical Enrichment Analysis and Predictive Modeling*, we evaluate the effect of database choice on several pathway enrichment methods and statistical modeling techniques by comparing the results obtained by three major pathway databases, as well as a merged representation of the three.

In this work, we first highlighted the rationales researchers often use to select a particular database for the analysis and interpretation of *omics* data, noting that these are often subjective, such as database popularity or previous experience. In some cases, tools may cater to a specific pathway format, and implicitly restrict an analysis to a particular database. Consequently, some databases are vastly over-represented in the literature than others, though hundreds of pathway resources are available [53]. Here, we focused on three major pathway databases (i.e., KEGG [36], Reactome [62] and WikiPathways [63]), given that all three are highly-cited, open-sourced and well-established. We also retrieved content from these primary databases from the integrative resource, MSigDB [151], to observe any effects that could be

attributed to outdated pathways stored in this meta-database. Furthermore, we generated MPath, an integrative resource containing the set union of KEGG, Reactome and WikiPathways, in which genesets and interactions of any pathway found in all three databases were merged. Mappings between pathways from the databases were established and retrieved from ComPath [65]. We hypothesized that if a pathway from one database is significantly enriched in an analysis, its equivalent representation in another database should presumably be significantly enriched as well. Thus, we aimed to systematically compare the results yielded by one resource over another.

We designed a series of experiments employing RNA-seq gene expression data from The Cancer Genome Atlas (TCGA) [152] for breast, kidney, liver, prostate and ovarian cancer. We subsequently designed a benchmarking schema using multiple enrichment methods to empirically test the use of one database over another on results. We chose methods that represented each of three generations of pathway enrichment methods, including (i) Fisher’s exact test [153] for over-representation analysis (ORA), (ii) GSEA [84] as a functional class scoring (FCS) method and, (iii) SPIA [154] as a pathway topology (PT)-based method, as well as ssGSEA [155], a single-sample FCS enrichment method. Finally, we benchmarked the performance of different pathway resources with regard to multiple machine learning tasks, specifically, the prediction of tumor vs. normal samples, tumor subtypes and overall survival.

We found that choosing one database over another has tremendous influence on the results of enrichment analysis, given their minimum overlap in terms of equivalent pathways [66]. We observed that different databases can yield disparate results on enrichment analysis and statistical modeling, even when the same pathway is represented in different databases, albeit with slightly altered gene sets and/or pathway topologies. For instance, we found that a pathway in one database may be significantly enriched for a particular dataset, but its equivalent representation in another database may not be. Similarly, a pathway could be over-expressed in results for an analysis conducted on one database, and under-expressed in results for another. Finally, we found that integrative resources which combine databases or relying upon multiple databases as opposed to a single one can be key to finding both biologically meaningful and consistent results. Ultimately, this benchmark study represents the first major attempt to shed light on the importance of database selection on pathway analysis and take a step towards more meaningful results.

Authors' contributions

Sarah Mubeen and Daniel Domingo Fernández conducted the main analysis and implemented the Python package. Holger Fröhlich supervised methodological aspects of the analysis. Charles Tapley Hoyt and André Gemünd assisted technically in the analysis of the results. Sarah Mubeen, Holger Fröhlich, Charles Tapley Hoyt, Martin Hofmann Apitius, and Daniel Domingo Fernández wrote the paper. Daniel Domingo Fernández conceived and designed the study.

2.3 DecoPath: a web application for decoding pathway enrichment analysis

This section presents the following publication (see **Appendix A.3**):

Sarah Mubeen, Vinay Srinivas Bharadhwaj, Alpha Tom Kodamullil, Yojana Gadiya, Martin Hofmann Apitius and Daniel Domingo Fernández (2021). DecoPath: A Web Application for Decoding Pathway Enrichment Analysis. *NAR Genomics and Bioinformatics*, 3(3): lqab087.

Summary

In the publication described in Section 2.1, we reviewed the major factors that the results of an enrichment analysis can hinge upon. We found that while several major comparative studies on various aspects of enrichment analysis have been conducted, such as the enrichment method, an oft-neglected yet crucial factor has remained the choice of pathway database. Then, in the work described in the preceding section (2.2), we identified significant effects on enrichment results owing to this factor. Until this publication, a comprehensive benchmark with objective evaluations on the impact of this factor was lacking, although some database selection guidelines could be found.

The findings from our previous benchmark study prompted us to develop a platform which allows users to easily identify differences in the results of enrichment analysis arising from the use of different resources. In the publication, *DecoPath: A Web Application for Decoding Pathway Enrichment Analysis*, we describe a novel web application, DecoPath, that can be used to identify exactly where differences lie when using one pathway database over another by comparing the results generated across databases. DecoPath allows researchers to conduct enrichment analysis using multiple enrichment methods and individual pathway databases as well as a merged representation. Built-in features, such as interactive visualizations, assist users in interpreting and gauging the reproducibility of their results.

Using the DecoPath ecosystem, researchers can run an enrichment analysis using the over representation analysis (ORA) or gene set enrichment analysis

(GSEA) ([84]) methods. Alternatively, researchers can upload the results of an analysis that have been performed on similar enrichment methods. DecoPath includes four major pathway databases (KEGG [36], PathBank [64], Reactome [62] and WikiPathways [63]) for pathway gene sets for which pathway mappings have already been established such that any equivalent pathway across the databases is noted as such. Once users have run an analysis to obtain results or alternatively, uploaded the results of an analysis, these results can be explored through custom visualizations from large-scale down to fine granular levels.

For a global overview of pathway analysis results, we created a pathway hierarchy in which pathways are organized into major categories (e.g., metabolism, immune system, signalling and disease pathways). The hierarchy itself is a directed acyclic graph with a maximum depth of four and contains pathways with either *is part of* or *equivalent to* relations types. In total, the hierarchy comprises of 644 pathways from the KEGG [36], PathBank [64], Reactome [62] and WikiPathways [63] databases. The pathway hierarchy can be viewed in the interactive hierarchical view visualization of equivalent pathways across databases. Here, researchers can visualize and explore nested pathways across multiple databases that either are or are not interesting to the phenotype under investigation.

At an intermediate level, individual pathways which are equivalent across database can be assessed to reveal the degree of consensus and/or discrepancies of the results of enrichment analysis across databases at the pathway level in the consensus page visualization. This visualization displays both the normalized enrichment score and whether the pathway is significantly enriched according to a user-defined significance cut-off for GSEA and exclusively the latter for ORA. In some cases, we found that an equivalent pathway across databases was significant in one database but not in another. Users can explore why that may be the case by conducting a gene-level analysis in parallel. At this deepest level of analysis, the contribution of individual genes within a given pathway can be studied to further facilitate the interpretation of results. This interactive visualization illustrates the overlap of genes for equivalent pathways to identify which genes are responsible for any contradictions observed in the results of enrichment analysis.

In conclusion, choosing one database over another can impact the results of enrichment analysis. By comparing the results obtained with multiple pathway databases, researchers acquire a broader, more comprehensive overview of the phenotype under study than if one were to rely upon a single database.

By investigating differences at the level of individual genes, one can also observe how alternative pathway definitions can determine whether or not a pathway is considered significantly enriched. Furthermore, such a gene-level analysis can also help to identify heavily annotated genes that contribute to non-specific enrichment results. One possible solution to the problem of non-specific results lies in drawing pathway boundaries such that these boundaries reduce redundancy across gene sets. In the following chapter, we introduce other modes of biological networks derived from data that can help guide functional network definitions instead of relying upon arbitrary pathway declarations.

Authors' contributions

Sarah Mubeen implemented the web application and analyzed the data with help from Vinay Srinivas Bharadhwaj and Daniel Domingo Fernández. Yojana Gadiya, Sarah Mubeen, and Daniel Domingo Fernández curated the pathway mappings. Sarah Mubeen and Daniel Domingo Fernández wrote the manuscript. Daniel Domingo Fernández conceived and designed the study.

CHAPTER 3

Revealing context-specific expression patterns through integrated biological networks

In this chapter, we present two publications which collectively investigate the expression patterns of genes across several contexts and conditions by leveraging various types of biological networks.

3.1 Towards a global investigation of transcriptomic signatures through co-expression networks and pathway knowledge for the identification of disease mechanisms

This section presents the following publication (see [Appendix A.4](#)):

Rebeca Queiroz Figueiredo, Tamara Raschka, Alpha Tom Kodamullil, Martin Hofmann Apitius, **Sarah Mubeen**[†] and Daniel Domingo Fernández[†] (2021). Towards a global investigation of transcriptomic signatures through co-expression networks and pathway knowledge for the identification of disease mechanisms. *Nucleic acid research*, 49(14): 7939–7953.

[†] Joint last authors

Summary

In this section and the one that follows, we introduce a specific class of biological networks in which nodes are connected depending on the strength of their correlations. These networks, termed gene co-expression networks, are often used for the analysis of experimental datasets as they allow for the simultaneous analysis of thousands of genes and facilitate the deciphering of gene expression patterns across multiple conditions.

In the publication, *Towards a global investigation of transcriptomic signatures through co-expression networks and pathway knowledge for the identification of disease mechanisms*, we identified nearly 4,500 disease-specific transcriptomic datasets from ArrayExpress which were filtered to retain datasets for patient-level data as well as control samples. Datasets which investigated the same or similar disease were grouped together under a common label. This resulted in the categorization of 38,621 samples from 469 datasets to 63 distinct diseases and one control group. Following batch correction to remove dataset specific effects and mapping of probes to genes, for patients samples for datasets categorized into each of the 63 diseases and their control samples, the transcriptomic data was then used to construct co-expression networks which represented the strongest gene-gene correlations in each of the different diseases and the normal group. These networks could be broadly grouped into ten distinct disease categories, including diseases of the cardiovascular, immune, gastrointestinal, respiratory, reproductive, nervous and musculoskeletal systems, as well as cancers, cognitive disorders, infectious diseases and others. Concurrently, we generated a human protein-protein interactome network containing nearly 200,000 interactions between approximately 8,600 protein-coding genes. Protein-protein interactions (PPIs) in this network were gathered from several resources, including multiple pathway databases.

The systematic nature of this work allowed us to investigate global expression patterns in hundreds of experimental datasets in a joint, knowledge- and data- driven manner. By constructing disease-specific co-expression networks from transcriptomic datasets and using a PPI network as a template, we were able to contextualize co-expression patterns that emerged across conditions. We conducted three different analyses of the co-expression networks, specifically at the node, edge and pathway levels, each of which was complemented by a parallel investigation of a PPI network. Firstly, we identified unique and common proteins across diseases in a node-level analysis and evaluated their

consistency against nodes from the PPI network as a proxy for their coverage in the scientific literature. In addition, we performed two subsequent analyses assuming that when a given pathway is relevant to a disease, the genes in the pathway will also tend to be strongly correlated in the disease-specific co-expression networks. Secondly, in an edge-level analysis, we studied the presence of common disease-specific correlations within the PPI network. Thirdly, in a pathway-level analysis, we explored the consensus and/or disagreements between connections in the disease specific co-expression networks and known PPIs in biological pathways.

Finally, having identified disease networks with a high similarity to specific pathways, we conducted a case study where we focused on a particular disease (i.e., schizophrenia) and an associated pathway (long-term potentiation (LTP)). We superimposed the co-expression network for schizophrenia with the LTP pathway, finding several edges in common. We also found that the vast majority of proteins in the LTP pathway were correlated in the co-expression network, illustrating that correlated proteins did tend towards involvement in the same biological process.

In conclusion, through this work, we demonstrate that when gene co-expression networks are superimposed with a protein protein interaction (PPI) network and pathway knowledge, one can connect the transcriptome with the proteome and contextualize gene co-expression patterns with prior, literature-based knowledge. Using a combined knowledge- and data- driven approach, we show how insights can be gained on common and unique mechanisms that underlie disease pathophysiology.

Authors' contributions

Daniel Domingo Fernández and Sarah Mubeen designed and supervised the study. Tamara Raschka implemented the pipeline to download, process and categorize the gene expression datasets. Rebeca Queiroz Figueiredo and Tamara Raschka generated the co-expression networks for each group. Sarah Mubeen and Daniel Domingo Fernández generated the interactome network. Rebeca Queiroz Figueiredo performed the analyses. Rebeca Queiroz Figueiredo, Tamara Raschka, Sarah Mubeen and Daniel Domingo Fernández interpreted the results. Rebeca Queiroz Figueiredo, Tamara Raschka, Sarah Mubeen and Daniel Domingo Fernández wrote the manuscript. Martin Hofmann Apitius conceived the original idea.

3.2 Elucidating gene expression patterns across multiple biological contexts through a large-scale investigation of transcriptomic datasets

This section presents the following publication (see [Appendix A.5](#)):

Rebeca Queiroz Figueiredo, Sara Díaz del Ser, Tamara Raschka, Martin Hofmann Apitius, Alpha Tom Kodamullil, **Sarah Mubeen**[†] and Daniel Domingo Fernández[†] (2022). Elucidating gene expression patterns across multiple biological contexts through a large-scale investigation of transcriptomic datasets. *BMC Bioinformatics*, 23(1).

[†] Joint last authors

Summary

Gene expression profiling enables the measurement of transcripts which are relevant to a certain condition or context, such as a cell. Patterns of genes expressed at the transcript level can be graphically organized into gene co-expression networks, where genes with correlated expression activity are connected. This sort of modelling is done under the premise that sets of genes involved in a specific biological process have similar patterns of expression.

In the previous section, we described a systematic approach for the study of disease-specific experimental datasets, identifying disease mechanisms through the integration of co-expression networks and pathway knowledge. Gene expression patterns are also observable in various other contexts, such as those responsible for normal physiology. Indeed, context-specific expression is often responsible for the diversity of functions and characterizations of cell types and tissues, each with their unique specializations. For instance, previous studies [156–158] have analyzed gene expression patterns at the cell and tissue level, finding significant cell and tissue type-specific expression signatures which aid in understanding human biology.

In the publication, *Elucidating gene expression patterns across multiple biological contexts through a large-scale investigation of transcriptomic datasets* we expanded the scope of our analysis to include multiple additional contexts beyond diseases, namely, cell types, tissues, and cell lines. Our work demonstrates a hybrid approach for the analysis of experimental data to discern which signatures of co-expression networks may be of particular significance to a certain cell or condition and which are constant across them.

Together with the preceding publication, we have presented a large-scale investigation of gene expression patterns across multiple biological contexts. While the previous section focused on characterizing transcriptomic signatures pertaining to disease-specific co-expression networks, the work done here was concerned with multiple contexts within a normal physiological state. In order to develop an overview of context-specific patterns that give rise to distinct biological processes, we collected over 600 experimental datasets and categorized them into nearly 100 sub-categories in one of the three studied contexts: cell types, tissues and cell lines. Examples of tissue types included as sub-categories were blood, kidney, liver, brain, lung and breast. Examples of cell types included dendritic cells, neurons, hepatocytes, stem cells, peripheral blood mononuclear cells, as well as its more specific cell types including monocytes, T cells, and lymphocytes. Finally, cell lines included those from breast cancer (e.g., MCF7 and SKBR3 cells), cervical cancer (HeLa cells), lung cancer (A549 and NCI-H1299 cells) and liver cancer (Huh7 and Hep G2 cells) amongst others.

Following the same procedure as the one presented in the preceding section (3.1), we retained only the strongest pairwise correlations between genes and constructed co-expression networks on which, analogous to section (3.1), we conducted a series of analyses at the node, pathway and network levels.

By once again leveraging a human protein-protein interaction network as a referential template, node level analyses identified nodes which were most common across contexts and their sub-categories and those which were already described in the literature. Pathway and network based analyses relied upon the PPI network to map observed correlations within the experimental data to pathway knowledge embedded in the PPI network. In the network-based analyses, we observed that the strongest correlations tended to correspond with PPIs more than expected by chance. We thus envision researchers can generate novel hypotheses of additional interactions in the PPI network from the gene-gene correlations present in the co-expression networks. In a similar manner, we also posit that the systematic overlay of pathways, context-

specific PPIs and co-expression networks generated for different contexts can help to re-define pathways by incorporating transcriptomic measurements. Another major findings of this work was that we were able to highlight that co-expression networks for bulk tissue can be inadequate in characterizing underlying cellular composition, emphasizing the importance of single-cell studies. Finally, in order to explore the networks generated in this work, we have made our findings freely available in a web application, along with data and scripts.

Authors' contributions

Daniel Domingo Fernández and Sarah Mubeen conceived and designed the study. Rebeca Queiroz Figueiredo and Tamara Raschka processed the transcriptomic datasets. Rebeca Queiroz Figueiredo implemented the methodology and analyzed the results supervised by Sarah Mubeen and Daniel Domingo Fernández. Sara Díaz del Ser implemented the web application. Rebeca Queiroz Figueiredo, Sarah Mubeen, and Daniel Domingo Fernández wrote the manuscript. Alpha Tom Kodamullil and Martin Hofmann Apitius reviewed the manuscript.

CHAPTER 4

Network-based algorithms for biological applications

The representation and organization of biological data in a network form can not only be far more intuitive for visual understanding, it can also facilitate the use of network-based methods for applications in computational network biology. In this chapter, we present network-based algorithms which leverage the relationships encoded within biological networks for various biomedical applications, such as drug discovery and gene function prediction.

4.1 MultiPaths: a Python framework for analyzing multi-layer biological networks using diffusion algorithms

This section presents the following publication (see **Appendix A.6**):

Josep Marín Llaó, Sarah Mubeen, Alexandre Perera Lluna, Martin Hofmann Apitius, Sergio Picart Armada, and Daniel Domingo Fernández. (2021). MultiPaths: a Python framework for analyzing multi-layer biological networks using diffusion algorithms. *Bioinformatics*, 37(1): 137-139.

Summary

Biological networks, such as pathways and biomedical knowledge graphs, serve as powerful paradigms to integrate and explain *omics* data. This is especially the case when these graphs are operationalized for algorithmic utility for various applications in biomedicine, leveraging algorithms used in domains as diverse as social networks, communication networks and networks of neural connections within the brain, amongst others. One particular algorithm, label propagation or network diffusion, is especially distinguished by its capacity to account for global network structure, finding uses in various biological applications, as described earlier in Section 1.3.3.

This algorithm relies upon the principle that genes within close proximity in a network tend to share common characteristics, such as their involvement in a particular biological process [110]. Under this assumption, network diffusion uses biological networks as powerful yet simplistic computational models for abstracting molecular interactions and associations. When the nodes of the network are superimposed with some prior knowledge or abstract label, a diffusion algorithm can diffuse or propagate a signal to neighbouring nodes, which can be inferred if they are unlabelled. Though numerous algorithms for diffusion exist, software to enable researchers to implement these algorithms are limited by the number of diffusion algorithms and biological databases they provide. In order to address these limitations, we have developed a framework to conduct network diffusion on biological networks.

The publication, *MultiPaths: a Python framework for analyzing multi-layer biological networks using diffusion algorithms*, presents MultiPaths, an ecosystem consisting of two Python packages for the analysis of biological networks. The first of the two packages, DiffuPy, implements several diffusion algorithms which are applicable to any generic network. Scoring schemes for propagating a label vector on a network are determined by a graph kernel which defines how the propagation behaves and spreads, how the input labels are codified and possible subsequent statistical normalization. The selection of one kernel or the use of one codification schema over another can lead to differences in results. Thus, DiffuPy provides a collection of five graph kernels, including the regularized Laplacian kernel, a matrix representation of a graph commonly used for diffusion [110]. Additionally, different diffusion methods can differ in how they codify positive (e.g., up-regulated entity), negative (e.g., down-regulated entity) and unlabelled (e.g., unmeasured entity) entities.

The second package offered by MultiPaths, DiffuPath, connects these algorithms to several multi-modal biological networks, thus, facilitating their utility in the scientific community. Specifically, the generic label propagation algorithms from DiffuPy can be applied to several biological networks we have made available in DiffuPath. These include biological networks encoded in various formats, including Simple Interaction Format (SIF) and Biological Expression Language (BEL) [70]. We provide three pathway databases (i.e., (KEGG [36], Reactome [62] and WikiPathways [63]) as well biological networks from several additional databases. These include disease-disease associations [128], DrugBank [94] for drug-target interactions, HSDN [159] for associations between diseases and symptoms, miRTarBase [160] for Interactions between miRNA and their targets, SIDER [161] for associations between drugs and side effects and Gene Ontology [95] for a hierarchy of tens of thousands of biological processes. Moreover, we created predefined collections so users can download pre-calculated kernels for sets of networks that represent integrated biological databases. These include an integrated representation of the three pathway databases, encompassing *-omics* modalities and biological processes/pathways [66], the merged representation and DrugBank [94], encompassing *-omics* modalities and biological processes/pathways with a strong focus on drug/chemical interactions and the merged representation and MirTarBase [160], encompassing *-omics* modalities and biological processes/pathways enriched with miRNAs.

Finally, this work has outlined several case scenarios conducted on multiple pathway networks, including the merged network representation, using multi-omics datasets. Specifically, we found that the merged multimodal network resulted in greater coverage of entities when compared to a network from any single resource, leading to improved performance metrics for diffusion algorithms in correctly identifying genes, metabolites and miRNAs. The results of these case scenarios highlight the following: i) the capacity of networks to accommodate heterogeneous data modalities, and ii) larger and more comprehensive networks generated from the combination of multiple resources yield better results compared with networks derived from individual ones.

Authors' contributions

Josep Marín Llaó implemented the methodology with assistance from Daniel Domingo Fernández and Sarah Mubeen. Sarah Mubeen created the networks

and network collections. Daniel Domingo Fernández, Josep Marín Llaó, Sergio Picart Armada and Sarah Mubeen performed the formal analyses. Sarah Mubeen, Daniel Domingo Fernández, Sergio Picart Armada and Josep Marín Llaó wrote the manuscript. Daniel Domingo Fernández conceived, designed and supervised the study.

4.2 Drug2ways: reasoning over causal paths in biological networks for drug discovery

This section presents the following publication (see **Appendix A.7**):

Daniel Rivas Barragan, **Sarah Mubeen**, Francesc Guim Bernat, Martin Hofmann Apitius, and Daniel Domingo Fernández (2020). Drug2ways: Reasoning over causal paths in biological networks for drug discovery. *PLOS Computational Biology*, 16(12): e1008464.

Summary

In the previous section, we introduced a class of algorithms which simulate the flow or diffusion of information through a partially labelled network to make inferences on unlabelled nodes for numerous applications, such as protein function prediction and drug target characterization. Diffusion algorithms heavily rely on the premise of network proximity, taking into account direct as well as distant neighbours. These algorithms are particularly powerful as they consider all possible paths between nodes within a network. Given the efficacy of this approach, in this section, we explored the prospect of leveraging the ensemble of paths between nodes for the especially challenging task of drug discovery, where both the high cost and attrition rate of the traditional approach to drug development impede the approval of novel as well as existing drugs for indications.

Oftentimes, given the small-world property of most biological networks and the computational complexity associated with exploring distant paths, pathfinding approaches in network-based drug discovery tend to be limited to

calculating cost-effective paths, such as shortest paths, between a drug and disease or disease phenotype. As outlined in Section 1.3.3, recent approaches in biomedicine have tended to focus on these shortest paths within a network, or drug discovery within discrete neighbourhoods of disease module sub-graphs [162]. However, therapeutic targets of a drug may not necessarily be located exclusively within close proximity to nodes which are relevant to a disease concept (e.g., disease-relevant gene). Rather, the ensemble of paths between a drug target and disease-relevant node may determine whether a drug could potentially be therapeutic for a disease.

In the publication *Drug2ways: Reasoning over causal paths in biological networks for drug discovery*, we introduce drug2ways, which, to our knowledge, is the first algorithm to traverse through all possible paths between pairs of nodes within a network for applications in drug discovery and drug repurposing.

Leveraging multi-modal causal networks, the drug2ways algorithm traverses along all causal paths between drug and disease nodes below a maximum length. A maximum length constraint was applied on the paths as exceedingly long paths and cycles can make the problem of calculating all paths intractable. We next defined the cumulative effect of a given drug on a disease-relevant node as the product of the effect of each individual intermediate node. The effect could be +1 or -1 depending on whether an entity activated or inhibited its neighbouring entity along the directed, causal path. Finally, drug-disease pairs whose ensemble of paths resulted in the reversal of the disease phenotype are then proposed as potential candidates. For instance, if the cumulative effect of a drug on a disease-relevant node was that of inhibition as a greater proportion of paths between the nodes below a given length were inhibitory, then the drug would be proposed by the algorithm as a potential therapeutic candidate for the disease and promising for further investigation.

To demonstrate drug2ways, we applied the algorithm on two distinct causal networks containing directed relationships between drugs, proteins, diseases and phenotypes. By obtaining a list of drug-disease pairs in clinical trials and mapping these to pairs within the two networks, we sought to validate our approach by testing drug2ways' ability to recover drug-disease pairs investigated in clinical trials among the top-ranked pairs proposed by the algorithm. The drug-disease pairs prioritized by drug2ways contained a large proportion of clinically investigated pairs and significantly outperformed both baseline and permuted networks, demonstrating the utility of the algorithm for applications in drug discovery and repurposing. Moreover, we presented

case scenarios in which the algorithm was also used to propose drug candidates that could simultaneously be used to optimize multiple targets (i.e., one drug with several disease and/or phenotypic targets), as well as drug candidates that could be used together in combination therapies. Finally, drug2ways is efficiently implemented in Python and freely available to the scientific community as a software package.

Authors' contributions

Daniel Rivas Barragan and Daniel Domingo Fernández implemented the methodology. Sarah Mubeen and Daniel Domingo Fernández generated the knowledge graphs for validation. Sarah Mubeen, Daniel Domingo Fernández, and Daniel Rivas Barragan performed the formal analyses. Sarah Mubeen, Daniel Domingo Fernández and Daniel Rivas Barragan wrote the manuscript. Daniel Domingo Fernández conceived, designed and supervised the study.

CHAPTER 5

Conclusion and outlook

In recent years, advanced technologies have led to the accumulation of vast quantities of biological data, in turn culminating in the paradoxical challenge of trying to ascertain how exactly each of these individual measured components fit together. By itself, this data is far too complex to interpret and meaningful patterns are essentially hidden from plain view. Moreover, many entities are engaged in complex interactions across biological scales that are overlooked in single *omic* experiments.

This dissertation has endeavoured to build a comprehensive picture of the mechanisms underlying normal and aberrant biological functioning by focusing on networks of interacting molecules across scales as the fundamental unit of study in lieu of individual molecules. Firstly, this has entailed bringing to light problems associated with pathway analysis for the interpretation of biological data and providing solutions to tackle these problems (Chapter 2). Moreover, we have integrated interaction data dispersed across heterogeneous resources towards building more complete biological networks which we use in several applications presented in this dissertation. Secondly, we have complemented the analysis of gene expression data with protein-protein interaction networks as a potential avenue to generate hypotheses of novel, context-specific links and to re-define pathway boundaries by leveraging transcriptomic measurements (Chapter 3). Thirdly, we operationalize biological networks for label propagation algorithms and for a novel pathfinding algorithm we have implemented for biomedical applications (Chapter 4).

The publications presented in Chapter 2 were primarily concerned with pathway analysis techniques that leverage pathway knowledge for the inter-

pretation of high throughput data. Here, we demonstrated that despite the value of pathway enrichment analysis, various avenues to conduct such an experiment can lead to substantial variability in results. This variability can stem from various factors, such as interchangeable modular components, the choice of pathway database or gene set collection, as well as distinct features of different experimental datasets. These and other factors were the subject of the review presented in Section 2.1. We especially focused on, i) supporting researchers in designing an experimental setup for this oft-used analysis which best reflects the research question at hand and, ii) advising researchers to bear in mind that modifications to the analysis can lead to varying results and interpretations, some of which may be inconsequential at best, and misleading at worst.

Then, in Section 2.2 and 2.3, we identified and tackled the problem of one major factor that contributes to variable results in pathway enrichment analysis, specifically, pathway database choice. In a benchmark study, we illustrated the critical importance of the careful consideration of this factor, demonstrating how different databases can yield disparate results. We also took a major step in consolidating several pathway databases such that researchers can simultaneously conduct multiple pathway analyses with different gene set collections and evaluate how results compare using the DecoPath web application we developed. Finally, we built an ontology which maps several pathway databases together, making it possible to investigate how alternative definitions of the same pathway can lead to differing results.

While the association of genes and pathways to a phenotype is a valuable insight, this sort of analysis serves as just one component of a much broader investigation of experimental data. In Chapter 3, we presented two publications aimed at contextualizing transcriptional patterns through a large-scale investigation of context-specific datasets. By charting gene expression patterns, we aimed at better understanding the pathophysiological mechanisms that lead to diseases or elucidate patterns which vary across different contexts. Specifically, publications presented in Sections 3.1 and 3.2 have explored four biological contexts: i) diseases, ii) tissues, iii) cell types, and iv) cell lines.

Across each of these contexts, we sought to divulge which patterns are uniquely characteristic to a given context, and which are recurrent across them. Furthermore, we sought to determine the degree to which edges in context-specific gene co-expression networks overlapped with known interactions in pathway and interaction databases. This was a challenging task given the constraints in overlaying these two disparate network types, and given that

up to 55% of protein interactions in PPI databases can be context-specific or transient [49]. Nonetheless, we were still able to highlight pathways that were highly similar to relevant context-specific networks. Finally, by deconstructing the patterns of each independent network, we endeavoured to shed light on biases and confounds which can occur due to overlapping contexts in experimental datasets, such as different compositions of cell types in tissues or diseases.

Finally, in Chapter 4, we demonstrated how abstracting biological data as networks can make them amenable to network-based algorithms to ask questions such as, which genes with as of yet unknown or obscure functions could be associated with a given biological process, or which active modules or communities might be relevant to a particular disease. In Section 4.1, we introduced MultiPaths, a Python framework which provides implementations of several network diffusion algorithms for the analysis of multimodal biological networks. In line with the work presented in Section 2.2, case scenarios conducted in this publication revealed that larger and more comprehensive networks generated by consolidating multiple heterogeneous resources can be more robust for analyses than individual ones.

Next, in Section 4.2, we presented drug2ways, a novel pathfinding algorithm which explores all paths between pairs of nodes in multimodal networks for drug discovery. Here, we hypothesized that a network-based approach which considers the ensemble of paths between drug and disease-relevant nodes within a network is more intuitive and biologically meaningful than similar pathfinding approaches for drug discovery that leverage network proximity methods (e.g., shortest paths). Using drug-disease pairs in clinical trials, we were able to validate our approach, finding that the top drug-disease pairs prioritized by drug2ways included a large proportion of clinically investigated ones.

As biological networks have become more and more widely used to organize and formalize biological data, their infrastructure to enable algorithmic utility has also become more robust. Nonetheless, the inherent complexity of biological systems, the transient and context-specific nature of interactions, obstacles with regards to data availability and interoperability, and a static overview that is intended to explain what is fundamentally a dynamic system, are all major limitations which remain to be addressed.

5.1 Future outlook

We foresee several future directions for the work that has been presented in this dissertation. One of the most pressing matters in relying upon pathways and networks for biological insights and biomedical applications is grappling with what is largely an incomplete picture of multimodal biomolecular interactions. These gaps in knowledge hold ramifications for our ability to piece together the mechanisms governing the biological processes we seek to investigate. Thus, first and foremost, we envisage the elucidation of more complete interactomes, in parallel to greater interoperability across heterogeneous resources which house their interactions. Nonetheless, recent advances in network biology are progressively seeing the development of more holistic, detailed and multi-modal networks constructed from heterogeneous resources, such as the PrimeKG to support precision medicine [163]. We anticipate future work will also expand and increasingly rely upon mappings across pathway databases, while the mappings we have created (see Section 2.3) can lay the foundation for a larger and more inclusive pathway ontology.

Within the context of this work, the addition of further pathway databases and mappings can facilitate pathway analysis by enabling a much broader coverage of pathway knowledge. More complete interactomes can also increase the capabilities of network-based algorithms to generate better predictions given that gaps in knowledge limit our abilities to accurately model biological mechanisms. Similarly, knowledge of all biomolecular interactions is crucial for applications which benefit from complementing knowledge and data-driven approaches, for example overlaying multi-*omic* measurements with pathway knowledge or interaction networks for drug discovery. However, methodologies that integrate high dimensional data can be noisy and complex (e.g., a genomics experiment can generate over 1 terabyte (TB) of data in a single run [164]). Moreover, techniques that model such data, can be fraught with challenges that occur from seemingly finding meaningful patterns in the data, but which are instead confounding factors and basal levels of gene expression.

Despite these challenges, future work focused on more granular investigations of co-expression data, such as single cell experiments, and the expansion to other contexts (e.g., species), can further help to differentiate between recurring and unique patterns. Additional lines of work that are crucial to understanding cellular functioning are considerations for not only individual cells, but also their local neighbourhood context. In particular, this can entail

analyzing cell-cell interactions using the techniques presented in Chapter 3 that leverage PPI networks, pathway knowledge and data-driven co-expression networks, specifically from single-cell datasets. By taking gene expression measurements of secreted extracellular proteins and their membrane-bound receptor proteins, and overlaying these with PPI networks, sub-graphs of their local neighbourhoods that model intercellular signalling can be generated, especially for applications in medicinal biology. Finally, incorporating temporal dimensions using longitudinal data [165] and quantitative and dynamic modelling [166] represent some of the next major frontiers in network biology.

References

1. Silverbush, D. & Sharan, R. A systematic approach to orient the human protein-protein interaction network. *Nature communications* **10**, 1–9 (2019).
2. Hsiao, A. & Kuo, M. D. High-throughput biology in the postgenomic era. *Journal of Vascular and Interventional Radiology* **20**, S488–96 (2009).
3. Hasin, Y., Seldin, M. & Lusi, A. Multi-omics approaches to disease. *Genome Biology* **18**, 1–15 (2017).
4. Manzoni, C. *et al.* Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. *Briefings in bioinformatics* **19**, 286–302 (2018).
5. Lightbody, G. *et al.* Review of applications of high-throughput sequencing in personalized medicine: barriers and facilitators of future progress in research and clinical application. *Briefings in bioinformatics* **20**, 1795–1811 (2019).
6. Cai, Z., Poulos, R. C., Liu, J. & Zhong, Q. Machine learning for multi-omics data integration in cancer. *iScience*, 103798 (2022).
7. Kim, J., Woo, H. R. & Nam, H. G. Toward systems understanding of leaf senescence: an integrated multi-omics perspective on leaf senescence research. *Molecular plant* **9**, 813–825 (2016).
8. Heyn, H. & Esteller, M. DNA methylation profiling in the clinic: applications and challenges. *Nature Reviews Genetics* **13**, 679–692 (2012).

9. Grant, G. R., Manduchi, E. & Stoeckert Jr, C. J. Analysis and management of microarray gene expression data. *Current protocols in molecular biology* **77**, 19–6 (2007).
10. Li, X. & Wang, C.-Y. From bulk, single-cell to spatial RNA sequencing. *International Journal of Oral Science* **13**, 1–6 (2021).
11. Svensson, V., Vento-Tormo, R. & Teichmann, S. A. Exponential scaling of single-cell RNA-seq in the past decade. *Nature protocols* **13**, 599–604 (2018).
12. Karlsson, M. *et al.* A single-cell type transcriptomics map of human tissues. *Science Advances* **7**, eabh2169 (2021).
13. Ramazi, S. & Zahiri, J. Posttranslational modifications in proteins: resources, tools and prediction methods. *Database* **2021** (2021).
14. Johnson, C. H., Ivanisevic, J. & Siuzdak, G. Metabolomics: beyond biomarkers and towards mechanisms. *Nature reviews Molecular cell biology* **17**, 451–459 (2016).
15. Roberts, L. D., Souza, A. L., Gerszten, R. E. & Clish, C. B. Targeted metabolomics. *Current protocols in molecular biology* **98**, 30–2 (2012).
16. Sender, R., Fuchs, S. & Milo, R. Revised estimates for the number of human and bacteria cells in the body. *PLoS biology* **14**, e1002533 (2016).
17. Mohajeri, M. H. *et al.* The role of the microbiome for human health: from basic science to clinical applications. *European journal of nutrition* **57**, 1–14 (2018).
18. Ursell, L. K., Metcalf, J. L., Parfrey, L. W. & Knight, R. Defining the human microbiome. *Nutrition reviews* **70**, S38–S44 (2012).
19. Gillies, R. J., Kinahan, P. E. & Hricak, H. Radiomics: images are more than pictures, they are data. *Radiology* **278**, 563–577 (2016).
20. Kim, M. *et al.* DNA methylation as a biomarker for cardiovascular disease risk. *PloS one* **5**, e9692 (2010).

21. Baylin, S. B. *et al.* Aberrant patterns of DNA methylation, chromatin formation and gene expression in cancer. *Human molecular genetics* **10**, 687–692 (2001).
22. Govender, M. A., Brandenburg, J.-T., Fabian, J. & Ramsay, M. The Use of ‘Omics for Diagnosing and Predicting Progression of Chronic Kidney Disease: A Scoping Review. *Frontiers in genetics* **12** (2021).
23. Yugi, K., Kubota, H., Hatano, A. & Kuroda, S. Trans-omics: how to reconstruct biochemical networks across multiple ‘omic’ layers. *Trends in biotechnology* **34**, 276–290 (2016).
24. Karczewski, K. J. & Snyder, M. P. Integrative omics for health and disease. *Nature Reviews Genetics* **19**, 299–310 (2018).
25. Misra, B. B., Langefeld, C., Olivier, M. & Cox, L. A. Integrated omics: tools, advances and future approaches. *Journal of molecular endocrinology* **62**, R21–R45 (2019).
26. Esteban-Gil, A. *et al.* ColPortal, an integrative multiomic platform for analysing epigenetic interactions in colorectal cancer. *Scientific data* **6**, 1–14 (2019).
27. Conesa, A. & Beck, S. Making multi-omics data accessible to researchers. *Scientific data* **6**, 1–4 (2019).
28. Gómez-Romero, L., López-Reyes, K. & Hernández-Lemus, E. The large scale structure of human metabolism reveals resilience via extensive signaling crosstalk. *Frontiers in physiology* **11**, 1667 (2020).
29. Sonawane, A. R., Weiss, S. T., Glass, K. & Sharma, A. Network medicine in the age of biomedical big data. *Frontiers in Genetics* **10**, 294 (2019).
30. Hawe, J. S., Theis, F. J. & Heinig, M. Inferring interaction networks from multi-omics data. *Frontiers in genetics* **10**, 535 (2019).
31. Szklarczyk, D. *et al.* The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic acids research* **49**, D605–D612 (2021).

32. Oughtred, R. *et al.* The BioGRID interaction database: 2019 update. *Nucleic acids research* **47**, D529–D541 (2019).
33. Alonso-López, D. *et al.* APID database: redefining protein–protein interaction experimental evidences and binary interactomes. *Database* **2019** (2019).
34. Hammal, F., de Langen, P., Bergon, A., Lopez, F. & Ballester, B. ReMap 2022: a database of Human, Mouse, Drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments. *Nucleic acids research* **50**, D316–D325 (2022).
35. Han, H. *et al.* TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic acids research* **46**, D380–D386 (2018).
36. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* **28**, 27–30 (2000).
37. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic acids research* **42**, D459–D471 (2014).
38. Bairoch, A. The ENZYME database in 2000. *Nucleic acids research* **28**, 304–305 (2000).
39. Chang, A. *et al.* BRENDA, the ELIXIR core data resource in 2021: new developments and updates. *Nucleic Acids Research* **49**, D498–D508 (2021).
40. Li, S. & Shui, W. Systematic mapping of protein–metabolite interactions with mass spectrometry-based techniques. *Current opinion in biotechnology* **64**, 24–31 (2020).
41. Diether, M., Nikolaev, Y., Allain, F. H. & Sauer, U. Systematic mapping of protein-metabolite interactions in central metabolism of *Escherichia coli*. *Molecular systems biology* **15**, e9008 (2019).
42. Lee, T. I. & Young, R. A. Transcriptional regulation and its misregulation in disease. *Cell* **152**, 1237–1251 (2013).

43. Lambert, S. A. *et al.* The human transcription factors. *Cell* **172**, 650–665 (2018).
44. Bushweller, J. H. Targeting transcription factors in cancer—from undruggable to reality. *Nature Reviews Cancer* **19**, 611–624 (2019).
45. Sanda, T. *et al.* Core transcriptional regulatory circuit controlled by the TAL1 complex in human T cell acute lymphoblastic leukemia. *Cancer cell* **22**, 209–221 (2012).
46. Hu, S. *et al.* Profiling the human protein-DNA interactome reveals ERK2 as a transcriptional repressor of interferon signaling. *Cell* **139**, 610–622 (2009).
47. Lin, Y. *et al.* RNAInter in 2020: RNA interactome repository with increased coverage and annotation. *Nucleic acids research* **48**, D189–D197 (2020).
48. Lapointe, C. P., Wilinski, D., Saunders, H. A. & Wickens, M. Protein-RNA networks revealed through covalent RNA marks. *Nature methods* **12**, 1163–1170 (2015).
49. Stacey, R. G., Skinnider, M. A., Chik, J. H. & Foster, L. J. Context-specific interactions in literature-curated protein interaction databases. *BMC Genomics* **19**, 1–10 (2018).
50. Institute, N. H. G. R. *Biological pathways* <https://www.genome.gov/27530687/biological-pathways-fact-sheet/>.
51. Joshi-Tope, G. *et al.* Reactome: a knowledgebase of biological pathways. *Nucleic acids research* **33**, D428–D432 (2005).
52. Chakazul. *Metabolic Metro Map* https://en.wikipedia.org/wiki/Metabolic_pathway.
53. Bader, G., Cary, M. & Sander, C. Pathguide: a pathway resource list. *Nucleic acids research* **34**, D504–D506 (2006).
54. Keseler, I. M. *et al.* The EcoCyc database: reflecting new knowledge about Escherichia coli K-12. *Nucleic acids research* **45**, D543–D550 (2017).

55. Naithani, S. *et al.* Plant Reactome: a knowledgebase and resource for comparative pathway analysis. *Nucleic acids research* **48**, D1093–D1103 (2020).
56. Domingo-Fernández, D. *et al.* Multimodal mechanistic signatures for neurodegenerative diseases (NeuroMMSig): a web server for mechanism enrichment. *Bioinformatics* **33**, 3679–3681 (2017).
57. Mizuno, S. *et al.* AlzPathway: a comprehensive map of signaling pathways of Alzheimer’s disease. *BMC systems biology* **6**, 1–10 (2012).
58. Kuperstein, I. *et al.* Atlas of Cancer Signalling Network: a systems biology resource for integrative analysis of cancer data with Google Maps. *Oncogenesis* **4**, e160–e160 (2015).
59. Perfetto, L. *et al.* SIGNOR: a database of causal relationships between biological entities. *Nucleic acids research* **44**, D548–D554 (2016).
60. Kandasamy, K. *et al.* NetPath: a public resource of curated signal transduction pathways. *Genome biology* **11**, 1–9 (2010).
61. Yang, T.-H., Wang, C.-C., Wang, Y.-C. & Wu, W.-S. YTRP: a repository for yeast transcriptional regulatory pathways. *Database* **2014** (2014).
62. Jassal, B. *et al.* The reactome pathway knowledgebase. *Nucleic acids research* **48**, D498–D503 (2020).
63. Martens, M. *et al.* WikiPathways: connecting communities. *Nucleic acids research* **49**, D613–D621 (2021).
64. Wishart, D. S. *et al.* PathBank: a comprehensive pathway database for model organisms. *Nucleic acids research* **48**, D470–D478 (2020).
65. Domingo-Fernández, D., Hoyt, C. T., Bobis-Álvarez, C., Marín-Llaó, J. & Hofmann-Apitius, M. ComPath: an ecosystem for exploring, analyzing, and curating mappings across pathway databases. *NPJ systems biology and applications* **4**, 1–8 (2018).
66. Domingo-Fernández, D., Mubeen, S., Marín-Llaó, J., Hoyt, C. T. & Hofmann-Apitius, M. PathMe: merging and exploring mechanistic pathway knowledge. *BMC bioinformatics* **20**, 1–12 (2019).

67. Belinky, F. *et al.* PathCards: multi-source consolidation of human biological pathways. *Database* **2015** (2015).
68. Demir, E. *et al.* The BioPAX community standard for pathway data sharing. *Nature biotechnology* **28**, 935–942 (2010).
69. Hucka, M. *et al.* The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* **19**, 524–531 (2003).
70. Slater, T. & Song, D. Saved by the BEL: ringing in a common language for the life sciences. *Drug Discovery World Fall* **2012**, 75–80 (2012).
71. Novère, N. L. *et al.* The systems biology graphical notation. *Nature biotechnology* **27**, 735–741 (2009).
72. Hermjakob, H. *et al.* The HUPO PSI’s molecular interaction format—a community standard for the representation of protein interaction data. *Nature biotechnology* **22**, 177–183 (2004).
73. Del Toro, N. *et al.* The IntAct database: efficient access to fine-grained molecular interaction data. *Nucleic acids research* **50**, D648–D653 (2022).
74. Oughtred, R. *et al.* The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Science* **30**, 187–200 (2021).
75. Calderone, A., Iannuccelli, M., Peluso, D. & Licata, L. Using the MINT database to search protein interactions. *Current Protocols in Bioinformatics* **69**, e93 (2020).
76. Miller, R. A. *et al.* Explicit interaction information from WikiPathways in RDF facilitates drug discovery in the Open PHACTS Discovery Platform. *F1000Research* **7** (2018).
77. Petri, V. *et al.* The pathway ontology—updates and applications. *Journal of biomedical semantics* **5**, 1–12 (2014).
78. Moreews, F., Simon, H., Siegel, A., Gondret, F. & Becker, E. PAX2GRAPHML: a python library for large-scale regulation network analysis using BioPAX. *Bioinformatics* **37**, 4889–4891 (2021).

79. Demir, E. *et al.* Using biological pathway data with paxtools. *PLoS computational biology* **9**, e1003194 (2013).
80. Rodchenkov, I. *et al.* Pathway Commons 2019 Update: integration, analysis and exploration of pathway data. *Nucleic acids research* **48**, D489–D497 (2020).
81. Herwig, R., Hardt, C., Lienhard, M. & Kamburov, A. Analyzing and interpreting genome data at the network level with ConsensusPathDB. *Nature protocols* **11**, 1889–1907 (2016).
82. Reimand, J. *et al.* Pathway enrichment analysis and visualization of omics data using g: Profiler, GSEA, Cytoscape and EnrichmentMap. *Nature protocols* **14**, 482–517 (2019).
83. Maleki, F., Ovens, K., Hogan, D. J. & Kusalik, A. J. Gene set analysis: challenges, opportunities, and future research. *Frontiers in genetics*, 654 (2020).
84. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102**, 15545–15550 (2005).
85. Hung, J.-H., Yang, T.-H., Hu, Z., Weng, Z. & DeLisi, C. Gene set enrichment analysis: performance evaluation and usage guidelines. *Briefings in bioinformatics* **13**, 281–291 (2012).
86. Liu, C. *et al.* Computational network biology: data, models, and applications. *Physics Reports* **846**, 1–66 (2020).
87. Pavlopoulos, G. A. *et al.* Using graph theory to analyze biological networks. *BioData mining* **4**, 1–27 (2011).
88. Mason, O. & Verwoerd, M. Graph theory and networks in biology. *IET systems biology* **1**, 89–119 (2007).
89. Cormen, T. H., Leiserson, C. E., Rivest, R. L. & Stein, C. *Introduction to algorithms* (MIT press, 2022).
90. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics* **9**, 1–13 (2008).

91. Pavlopoulos, G. A. *et al.* Bipartite graphs in systems biology and medicine: a survey of methods and applications. *GigaScience* **7**, giy014 (2018).
92. Department of Computer Science, S. U. *Knowledge Graphs* https://web.stanford.edu/class/cs520/2020/notes/What_is_a_Knowledge_Graph.html.
93. Ehrlinger, L. & Wöß, W. Towards a definition of knowledge graphs. *SEMANTiCS (Posters, Demos, SuCCESS)* **48**, 2 (2016).
94. Wishart, D. S. *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic acids research* **46**, D1074–D1082 (2018).
95. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nature genetics* **25**, 25–29 (2000).
96. Nicholson, D. N. & Greene, C. S. Constructing knowledge graphs and their biomedical applications. *Computational and structural biotechnology journal* **18**, 1414–1428 (2020).
97. Morton, K. *et al.* ROBOKOP: an abstraction layer and user interface for knowledge graphs to support question answering. *Bioinformatics* **35**, 5382–5384 (2019).
98. Karim, M. R. *et al.* Drug-drug interaction prediction based on knowledge graph embeddings and convolutional-LSTM network in *Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics* (2019), 113–123.
99. Lin, X., Quan, Z., Wang, Z.-J., Ma, T. & Zeng, X. *KGNN: Knowledge Graph Neural Network for Drug-Drug Interaction Prediction*. in *IJCAI* **380** (2020), 2739–2745.
100. Celebi, R. *et al.* Evaluation of knowledge graph embedding approaches for drug-drug interaction prediction in realistic settings. *BMC bioinformatics* **20**, 1–14 (2019).
101. Yu, Y. *et al.* SumGNN: multi-typed drug interaction prediction via efficient knowledge graph summarization. *Bioinformatics* **37**, 2988–2995 (2021).

102. Dai, Y., Guo, C., Guo, W. & Eickhoff, C. Drug–drug interaction prediction with Wasserstein Adversarial Autoencoder-based knowledge graph embeddings. *Briefings in bioinformatics* **22**, bbaa256 (2021).
103. Abdelaziz, I., Fokoue, A., Hassanzadeh, O., Zhang, P. & Sadoghi, M. Large-scale structural and textual similarity-based mining of knowledge graph to predict drug–drug interactions. *Journal of Web Semantics* **44**, 104–117 (2017).
104. Mohamed, S. K., Nounu, A. & Nováček, V. *Drug target discovery using knowledge graph embeddings* in *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing* (2019), 11–18.
105. Bean, D. M. *et al.* Knowledge graph prediction of unknown adverse drug reactions and validation in electronic health records. *Scientific reports* **7**, 1–11 (2017).
106. Zitnik, M., Agrawal, M. & Leskovec, J. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics* **34**, i457–i466 (2018).
107. Dai, W. *et al.* Matrix factorization-based prediction of novel drug indications by integrating genomic space. *Computational and mathematical methods in medicine* **2015** (2015).
108. Youn, J., Rai, N. & Tagkopoulos, I. Knowledge integration and decision support for accelerated discovery of antibiotic resistance genes. *Nature Communications* **13**, 1–11 (2022).
109. Li, L. *et al.* Real-world data medical knowledge graph: construction and applications. *Artificial intelligence in medicine* **103**, 101817 (2020).
110. Cowen, L., Ideker, T., Raphael, B. J. & Sharan, R. Network propagation: a universal amplifier of genetic associations. *Nature Reviews Genetics* **18**, 551–562 (2017).
111. Domingo-Fernández, D. *et al.* Causal reasoning over knowledge graphs leveraging drug-perturbed and disease-specific transcriptomic signatures for drug discovery. *PLoS computational biology* **18**, e1009909 (2022).

112. Zhu, X. & Ghahramani, Z. Learning from labeled and unlabeled data with label propagation. *Technical Report CMU-CALD-02-107, Carnegie Mellon University* (2002).
113. Warde-Farley, D. *et al.* The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic acids research* **38**, W214–W220 (2010).
114. Leiserson, M. D. *et al.* Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature genetics* **47**, 106–114 (2015).
115. Wang, B. *et al.* Similarity network fusion for aggregating data types on a genomic scale. *Nature methods* **11**, 333–337 (2014).
116. Needham, M. & Hodler, A. E. *Graph algorithms: practical examples in Apache Spark and Neo4j* (O’Reilly Media, 2019).
117. Lee, T. & Yoon, Y. Drug repositioning using drug-disease vectors based on an integrated network. *BMC bioinformatics* **19**, 1–12 (2018).
118. Yu, H. *et al.* Prediction of drugs having opposite effects on disease genes in a directed network. *BMC systems biology* **10**, 17–25 (2016).
119. Isik, Z., Baldow, C., Cannistraci, C. V. & Schroeder, M. Drug target prioritization by perturbed gene expression and network information. *Scientific reports* **5**, 1–13 (2015).
120. Kotlyar, M., Fortney, K. & Jurisica, I. Network-based characterization of drug-regulated genes, drug targets, and toxicity. *Methods* **57**, 499–507 (2012).
121. Fortunato, S. & Hric, D. Community detection in networks: A user guide. *Physics reports* **659**, 1–44 (2016).
122. Girvan, M. & Newman, M. E. Community structure in social and biological networks. *Proceedings of the national academy of sciences* **99**, 7821–7826 (2002).
123. Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. From molecular to modular cell biology. *Nature* **402**, C47–C52 (1999).

124. Choobdar, S. *et al.* Assessment of network module identification across complex diseases. *Nature methods* **16**, 843–852 (2019).
125. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic acids research* **30**, 1575–1584 (2002).
126. Huttlin, E. L. *et al.* Architecture of the human interactome defines protein communities and disease networks. *Nature* **545**, 505–509 (2017).
127. Gustafsson, M. *et al.* Modules, networks and systems medicine for understanding disease and aiding diagnosis. *Genome medicine* **6**, 1–11 (2014).
128. Menche, J. *et al.* Uncovering disease-disease relationships through the incomplete interactome. *Science* **347**, 1257601 (2015).
129. Ghiassian, S. D., Menche, J. & Barabási, A.-L. A DIseAse MOdule Detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. *PLoS computational biology* **11**, e1004120 (2015).
130. Lü, L. & Zhou, T. Link prediction in complex networks: A survey. *Physica A: statistical mechanics and its applications* **390**, 1150–1170 (2011).
131. Kovács, I. A. *et al.* Network-based prediction of protein interactions. *Nature communications* **10**, 1–8 (2019).
132. Abbas, K. *et al.* Application of network link prediction in drug discovery. *BMC bioinformatics* **22**, 1–21 (2021).
133. Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C. & Collins, J. J. Next-generation machine learning for biological networks. *Cell* **173**, 1581–1592 (2018).
134. Hamilton, W. L., Ying, R. & Leskovec, J. Representation learning on graphs: Methods and applications. *IEEE Data Engineering Bulletin* **40**, 52–74 (2017).

135. Cai, H., Zheng, V. W. & Chang, K. C.-C. A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE Transactions on Knowledge and Data Engineering* **30**, 1616–1637 (2018).
136. Atz, K., Grisoni, F. & Schneider, G. Geometric deep learning on molecular representations. *Nature Machine Intelligence* **3**, 1023–1032 (2021).
137. Belkin, M. & Niyogi, P. Laplacian eigenmaps and spectral techniques for embedding and clustering. *Advances in neural information processing systems* **14** (2001).
138. Grover, A. & Leskovec, J. *node2vec: Scalable feature learning for networks* in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining* (2016), 855–864.
139. Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representations (ICLR)* (2017).
140. Nickel, M., Rosasco, L. & Poggio, T. *Holographic embeddings of knowledge graphs* in *Proceedings of the AAAI Conference on Artificial Intelligence* **30** (2016).
141. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J. & Yakhnenko, O. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems* **26** (2013).
142. Dong, Y., Chawla, N. V. & Swami, A. *metapath2vec: Scalable representation learning for heterogeneous networks* in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining* (2017), 135–144.
143. Su, C., Tong, J., Zhu, Y., Cui, P. & Wang, F. Network embedding in biomedical data science. *Briefings in bioinformatics* **21**, 182–197 (2020).
144. Kulmanov, M., Khan, M. A. & Hoehndorf, R. DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics* **34**, 660–668 (2018).

145. Ratajczak, F., Joblin, M., Ringsquandl, M. & Hildebrandt, M. Task-driven knowledge graph filtering improves prioritizing drugs for repurposing. *BMC bioinformatics* **23**, 1–19 (2022).
146. Peng, J., Guan, J. & Shang, X. Predicting Parkinson’s disease genes based on node2vec and autoencoder. *Frontiers in genetics* **10**, 226 (2019).
147. Luo, Y. *et al.* A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nature communications* **8**, 1–13 (2017).
148. Wang, S. *et al.* Identification of pathways associated with chemosensitivity through network embedding. *PLoS computational biology* **15**, e1006864 (2019).
149. Li, Y. & Yang, T. in *Guide to big data applications* 83–104 (Springer, 2018).
150. Balabin, H. *et al.* STonKGs: a sophisticated transformer trained on biomedical text and knowledge graphs. *Bioinformatics* **38**, 1648–1656 (2022).
151. Liberzon, A. *et al.* The molecular signatures database hallmark gene set collection. *Cell systems* **1**, 417–425 (2015).
152. Weinstein, J. N. *et al.* The cancer genome atlas pan-cancer analysis project. *Nature genetics* **45**, 1113–1120 (2013).
153. Fisher, R. A. in *Breakthroughs in statistics* 66–70 (Springer, 1992).
154. Tarca, A. L. *et al.* A novel signaling pathway impact analysis. *Bioinformatics* **25**, 75–82 (2009).
155. Barbie, D. A. *et al.* Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* **462**, 108–112 (2009).
156. Dobrin, R. *et al.* Multi-tissue coexpression networks reveal unexpected subnetworks associated with disease. *Genome biology* **10**, 1–3 (2009).
157. Pierson, E., Consortium, G., Koller, D., Battle, A. & Mostafavi, S. Sharing and specificity of co-expression networks across 35 human tissues. *PLoS computational biology* **11**, e1004220 (2015).

158. McKenzie, A. *et al.* Brain cell type specific gene expression and co-expression network architectures. *Scientific reports* **8**, 1–9 (2018).
159. Zhou, X., Menche, J., Barabási, A.-L. & Sharma, A. Human symptoms-disease network. *Nature communications* **5**, 1–10 (2014).
160. Huang, H.-Y. *et al.* miRTarBase 2020: updates to the experimentally validated microRNA–target interaction database. *Nucleic acids research* **48**, D148–D154 (2020).
161. Kuhn, M., Letunic, I., Jensen, L. J. & Bork, P. The SIDER database of drugs and side effects. *Nucleic acids research* **44**, D1075–D1079 (2016).
162. Cheng, F., Kovács, I. A. & Barabási, A.-L. Network-based prediction of drug combinations. *Nature communications* **10**, 1–11 (2019).
163. Chandak, P., Huang, K. & Zitnik, M. Building a knowledge graph to enable precision medicine. *bioRxiv* (2022).
164. Hasenauer, J., Banga, J. R., Martina-Perez, S., Sailem, H. & Baker, R. E. Mathematical modelling of high-throughput and high-content data. *Current Opinion in Systems Biology* **29**, 100405 (2022).
165. Bodein, A., Scott-Boyer, M.-P., Perin, O., Lê Cao, K.-A. & Droit, A. Interpretation of network-based integration from multi-omics longitudinal data. *Nucleic acids research* **50**, e27–e27 (2022).
166. Villaverde, A. F., Pathirana, D., Fröhlich, F., Hasenauer, J. & Banga, J. R. A protocol for dynamic model calibration. *Briefings in bioinformatics* **23**, bbab387 (2022).

APPENDIX A





Appendix

A.1 On the influence of several factors on pathway enrichment analysis

Reprinted with permission from “Mubeen S., Kodamullil A.T., Hofmann-Apitius M., and Domingo-Fernández D. (2022). On the influence of several factors on pathway enrichment analysis *Briefings in Bioinformatics*, 23(3)”.

Copyright © Mubeen, S., *et al.*, 2022.

On the influence of several factors on pathway enrichment analysis

Sarah Mubeen , Alpha Tom Kodamullil , Martin Hofmann-Apitius  and Daniel Domingo-Fernández 

Corresponding authors: Sarah Mubeen, Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Sankt Augustin 53757, Germany. Tel: +49 2241 14-4204; E-mail: sarah.mubeen@scai.fraunhofer.de; Daniel Domingo-Fernández, Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Sankt Augustin 53757, Germany. Tel: +49 2241 14-4036; E-mail: daniel.domingo.fernandez@scai.fraunhofer.de

Abstract

Pathway enrichment analysis has become a widely used knowledge-based approach for the interpretation of biomedical data. Its popularity has led to an explosion of both enrichment methods and pathway databases. While the elegance of pathway enrichment lies in its simplicity, multiple factors can impact the results of such an analysis, which may not be accounted for. Researchers may fail to give influential aspects their due, resorting instead to popular methods and gene set collections, or default settings. Despite ongoing efforts to establish set guidelines, meaningful results are still hampered by a lack of consensus or gold standards around how enrichment analysis should be conducted. Nonetheless, such concerns have prompted a series of benchmark studies specifically focused on evaluating the influence of various factors on pathway enrichment results. In this review, we organize and summarize the findings of these benchmarks to provide a comprehensive overview on the influence of these factors. Our work covers a broad spectrum of factors, spanning from methodological assumptions to those related to prior biological knowledge, such as pathway definitions and database choice. In doing so, we aim to shed light on how these aspects can lead to insignificant, uninteresting or even contradictory results. Finally, we conclude the review by proposing future benchmarks as well as solutions to overcome some of the challenges, which originate from the outlined factors.

Keywords: pathway enrichment, gene set analysis, pathway database, omics data, benchmark, gene set collection

Introduction

Pathway enrichment analysis has become one of the foremost methods for the interpretation of biological data as it facilitates the reduction of high-dimensional information to just a handful of biological processes underlying specific phenotypes. Over the last decade, the popularity of pathway enrichment analysis has led to the development of numerous different methods that can be categorized into three generations: (i) over-representation analysis (ORA), (ii) functional class scoring (FCS) and (iii) pathway topology (PT)-based, each of which adds an increasing layer of complexity to the analysis [1]. ORA, the first of the three, refers to a class of methods designed to identify gene sets that share a larger number of genes in common with a list of differentially expressed genes (DEGs) than would be expected by chance. Given a list of DEGs, a gene set and their complements, a statistical test is conducted

to assess whether DEGs are over-represented in the gene set. Though simple to conduct, ORA methods rely upon arbitrary, and at times harsh, cutoffs to determine what constitutes a DEG. To remedy this problem, FCS methods test whether genes of a gene set have coordinated activity with the phenotype under study by using metrics to assign differential expression scores to each gene in the experiment. Genes are then ranked by their scores, which are subsequently used to calculate gene set scores and determine gene sets that are interesting in some statistically significant way. Finally, PT-based approaches build upon the latter class of methods and are characterized as additionally taking PT information into account, rather than solely relying upon gene sets, which lack interaction information. Thus, a formal distinction can be made between gene sets and pathways. Specifically, a gene set refers to a set of unranked genes which can be variously grouped, such

Sarah Mubeen is a doctoral student at the University of Bonn and is a research fellow at the Department of Bioinformatics at the Fraunhofer Institute for Scientific Computing and Algorithms (SCAI). Her research interests include pathway and network-based approaches for the interpretation of biomedical data.

Alpha Tom Kodamullil is the group lead of Applied Semantics at the Fraunhofer Institute SCAI. Her main research focuses on knowledge graphs around diseases and shared semantics, which lay the foundation of data and knowledge interoperability.

Martin Hofmann-Apitius is a professor at the Bonn-Aachen International Center for Information Technology at the University of Bonn and is the head of the Department of Bioinformatics at the Fraunhofer Institute SCAI. His research interests involve integrative semantics, natural language processing, data- and knowledge-integration in neurodegeneration research and longitudinal modeling of disease progression.

Daniel Domingo-Fernández is a research fellow at the Department of Bioinformatics at the Fraunhofer Institute SCAI and is a senior bioinformatician at Enveda Biosciences. His research interests lie in integrating a priori knowledge and biomedical data for drug discovery, predictive modeling and patient stratification.

Received: January 17, 2022. **Revised:** March 21, 2022. **Accepted:** March 30, 2022

© The Author(s) 2022. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

as by their membership within a biological pathway or chromosomal position, while a pathway refers to a set of genes as well as any pairwise interactions between them. While the simplicity and accessibility of enrichment methods have been the main drivers to their widespread adoption by the community, the broad pool of methods at hand and the lack of gold standards pose a challenge in evaluating the variability of enrichment results. Consequently, several guidelines have been published in recent years on recommendations for the experimental design of an enrichment analysis [2–4].

An analogous but more philosophical debate in the community pertains to the choice of pathway or gene set database. Its selection is arguably one of the most decisive factors influencing the results of enrichment analyses as it determines the possible gene sets that can be enriched (i.e. genes within a gene set are enriched in an examined list of genes). The number of public databases has continued to grow in the past years in parallel with novel enrichment methods. However, the list of the most widely used databases has not changed in the last decade as enrichment analyses are predominantly conducted exclusively on one of the following three databases: KEGG [5], Reactome [6] and Gene Ontology (GO) [7]. While this selected group of databases comes with several advantages (e.g. large coverage of biological processes and regular updates), definitions of what constitutes a given pathway or gene set may be arbitrarily drawn across databases.

At present, users are offered a wide spectrum of enrichment methods and databases when performing enrichment analyses. This poses a challenge when considering the numerous factors that play a role in results of enrichment analysis, which can lead to insignificant, irrelevant or even contradictory results. Thus, in recent years, several benchmark studies have been conducted to evaluate the effects of various aspects of pathway analysis for practical guidelines.

In this work, we review the findings of major benchmarks conducted on different factors that influence the results of pathway enrichment analysis (Figure 1). The goal of our paper is to both inform the broader community of researchers using pathway enrichment analysis of these factors and to summarize the findings of all the most recent benchmarks. Finally, we also discuss possible solutions to address these factors as well as other factors that have not yet been investigated but can be benchmarked in the future.

Comparative studies on enrichment methods

Given the popularity of pathway enrichment analysis, at least 70 different methods have been developed as well as hundreds of variants [8, 9] (see Xie *et al.* [10] for an exhaustive survey of methods and benchmarks). The implementations of these methods can differ based on a number of factors, such as the gene-level statistic

(e.g. t-test statistic and fold change), the gene set-level statistic (e.g. Kolmogorov–Smirnov (KS) statistic [11] and Wilcoxon rank sum test [12]), the formulation of the null and alternative hypotheses and the significance estimate. Many of the most commonly employed pathway enrichment methods have been compared in several major benchmarks and reviews. In this section, we outline the findings of 12 comprehensive comparative studies on enrichment methods (Table 1; for more details, see Supplementary Tables 1–3 available online at <https://academic.oup.com/bib>).

Metrics for method evaluation

A particular challenge in the design of comparative studies on enrichment methods is that in the absence of a comprehensive understanding of the complex biological processes involved across experimental conditions, results are often not verifiable beyond retrospective evaluations. That is to say, without a gold standard with which to compare the results produced by any given method, conclusive assessments are often difficult to make. Nonetheless, several techniques to compare methods are widely used, while benchmark datasets have also been proposed. Specifically, datasets used by benchmark studies reviewed herein have largely been real, experimental datasets investigating a particular phenotype (i.e. the object of study in the experiment). Following Tarca *et al.* [23], several studies [2, 3, 9, 13, 25] have selected evaluation datasets as those which correspond to a pathway or gene set from the chosen database (e.g. dataset investigating the breast cancer versus normal phenotype and the breast cancer pathway). Others [14, 16, 24] have focused on measuring consistency across methods by selecting various datasets that study the same phenotype. Finally, comparative studies [3, 13, 14, 18, 22] have also employed simulated datasets to benchmark methods as various features of the data can be tuned and the method can be studied under these known features of the data. In line with Tarca *et al.* [23], the majority of studies have evaluated the performance of an enrichment method on these datasets based on at least one of the following metrics: prioritization, specificity or sensitivity.

Prioritization is evaluated based on whether a target gene set that has been identified a priori as showing high relevance to a phenotype associated with the dataset under investigation is ranked near the top (e.g. the breast cancer pathway is expected to hold the topmost ranking for a dataset measuring transcriptomic differences between the breast cancer versus normal phenotypes). Specificity refers to the proportion of gene sets that are correctly identified by a method as true negatives; thus, methods with a high specificity will generate fewer false positives. Finally, of all the gene sets detected as significant by a given method, sensitivity measures the proportion of gene sets that are actually relevant to the phenotype associated with the dataset under study (i.e. true positives).

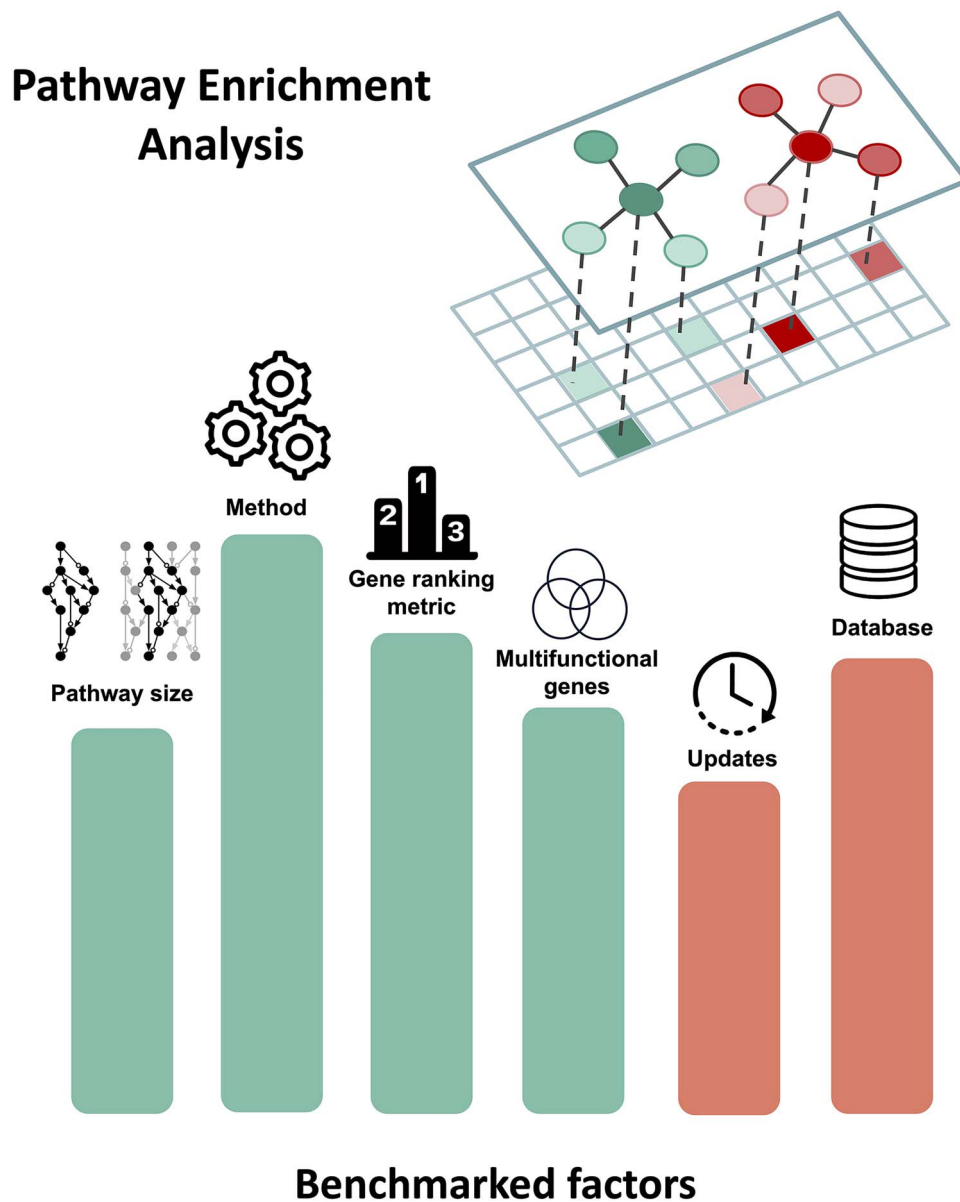


Figure 1. Illustration of major factors that influence the results of pathway enrichment analysis discussed in this review. The height and color of the bars are symbolic and do not correlate with importance. The two networks depicted above represent two biological pathways mapped to gene expression data (matrix below).

Of the various comparative studies done to date, the above-mentioned metrics have been among the most commonly used for the empirical evaluation of enrichment methods. Nonetheless, the metrics used and the methods benchmarked by an individual study can vary greatly, with the most popular methods, not surprisingly, studied the most frequently. Yet despite the numerous benchmark studies conducted thus far, a comprehensive and standardized assessment of the many enrichment methods available has yet to be performed. Moreover, of the benchmark studies that have attempted such an assessment, no specific method has been shown to yield consistent results across all evaluated settings. Nevertheless, trends do emerge regarding the individual performance of a method on a given metric (Supplementary Tables 4–6 available online

at <https://academic.oup.com/bib>). Thus, in the following, we report the trends observed across comparative studies for methods that consistently show superior performance on metrics in two or more studies without showing a poor performance on that same metric.

With regard to sensitivity, MRGSE [26], GlobalTest [27] and PLAGE [28] ranked highly in studies by Tarca *et al.* [23] and Zyla *et al.* [25] (Supplementary Table 4 available online at <https://academic.oup.com/bib>). However, high sensitivity may also imply a lower specificity. This was indeed observed for MRGSE and PLAGE, both of which reported a larger than expected number of false positives in at least one study, though also a good performance in prioritization (Supplementary Table 6 available online at <https://academic.oup.com/bib>). This is not surprising given that both methods have also been shown to report

Table 1. Comparative studies evaluating differences across enrichment methods

No.	Review	Methods tested	Datasets	Database (# of gene sets/pathways)	Types of evaluated methods
1	[13]	7	36	KEGG (116)	Topology- and non-topology-based methods
2	[2]	10	75	KEGG (323) and GO (4631)	ORA and FCS methods
3	[3]	7	118	KEGG (232)	Topology-based methods
4	[14]	6	20	KEGG (86)	Topology- and non-topology-based methods
5	[15]	9	3	KEGG (114)	Topology-based methods
6	[16]	13	6	GO gene set collection extracted from MSigDB [17] v6.1 (5917)	Widely used pathway enrichment methods
7	[18]	8	3	MSigDB v5.0 (10,295)	Widely used pathway enrichment methods
8	[9]	10	86	KEGG; 150 pathways for all methods except 130 for PathNet [19] and 186 for CePa [20, 21]	Topology- and non-topology-based methods
9	[22]	11	1	C2 collection from MSigDB v4.0 (4722)	Methods differing based on null hypothesis
10	[23]	16	42	KEGG (259) and Metacore™ (88)	ORA and FCS methods
11	[24]	5	6	KEGG (192)	ORA and FCS methods
12	[25]	7	38	KEGG (189)	ORA and FCS methods

In the third column, we report the number of enrichment methods compared in each study (see [Supplementary Tables 2 and 3](#), available online at <https://academic.oup.com/bib>, for details on the methods tested). Here, we would like to note that we differentiate between methods and tools/web applications based on Geistlinger *et al.* [2]. In the fourth column, we report the number of datasets each study performed comparisons on, all of which were experimental datasets except in [3, 13, 14, 18, 22], which included both experimental and simulated datasets. Finally, the fifth column reports the pathway databases used in each study while the number of pathways is shown between parentheses.

a majority of gene sets as significant [24, 25]. Similarly, classical statistical tests, including the KS test and the Wilcoxon rank sum test, were highly sensitive in Bayerlová *et al.* [13] and Nguyen *et al.* [9], though results were inconsistent regarding their specificity. Notably, of the above-mentioned methods, GlobalTest was the only investigated method to consistently demonstrate high sensitivity as well as high specificity in studies by Tarca *et al.* [23] and Zyla *et al.* [25].

In assessments of specificity, SPIA [29] and CAMERA [30] have shown high specificity in at least two studies ([Supplementary Table 5](#) available online at <https://academic.oup.com/bib>), though results have been mixed or poor with regard to sensitivity and target pathway prioritization. Furthermore, GSA [31], PADOG [32] and PathNet showed good results with regard to prioritization ([Supplementary Table 6](#) available online at <https://academic.oup.com/bib>) but mixed results for sensitivity and specificity. Finally, across all studies, GSEA [33] and ORA (or a variant) were the most investigated enrichment methods, with 8 of 12 comparative studies assessing either one or both of these methods ([Supplementary Table 3](#) available online at <https://academic.oup.com/bib>). Here, we observed that, although they were the most commonly used methods for enrichment analysis, results regarding their sensitivity, specificity and prioritization were altogether inconsistent ([Supplementary Tables 4–6](#) available online at <https://academic.oup.com/bib>).

Hypothesis testing and significance assessment

Much of the focus of comparative analyses on gene set analysis methods has been on the implications of alternative definitions of the null hypothesis. In their seminal work, Goeman and Bühlmann [34] characterized methods by the null hypothesis assumed in the statistical test. Enrichment methods, they assert, can be categorized as

being competitive methods if they test the competitive null hypothesis [i.e. those which assume that genes in a gene set are not differentially expressed with respect to their complement (typically the rest of the genes in the experiment)] or self-contained methods if they test the self-contained null hypothesis (i.e. those which assume that genes in a gene set are not differentially expressed across phenotypes). Choosing one category of methods over another can confer several advantages, which we explicate through a brief review of studies that have assessed the performance of methods, which differ based on this distinction.

Rahmatallah *et al.* [22] recapitulated earlier work [35–37], generally noting that the power of self-contained methods was greater than that of competitive ones ([Table 1](#); [Supplementary Tables 2 and 3](#) available online at <https://academic.oup.com/bib>). Self-contained methods were also more robust to sample size and heterogeneity, with these methods showing the highest sensitivity among all the ones they evaluated, even as the sample sizes decreased [22] ([Supplementary Table 7](#) available online at <https://academic.oup.com/bib>). Specifically, they found that ROAST [38] and SAM-GS [39] yielded the best performance on this metric.

Geistlinger *et al.* [2] noted that the proportions of gene sets reported as significant by methods differed based on the type of null hypothesis tested. Out of 10 investigated methods ([Supplementary Table 3](#) available online at <https://academic.oup.com/bib>), they found that the majority of self-contained ones, including GlobalTest, detected a larger fraction of gene sets as significant. In Zyla *et al.* [25], the self-contained methods GlobalTest and PLAGE also reported the largest number of gene sets as significant among all benchmarked methods ([Supplementary Table 3](#) available online at <https://academic.oup.com/bib>). In contrast to these findings, Wu and Lin [37] found that GlobalTest reported

fewer gene sets as significantly enriched in comparison with competitive methods.

Furthermore, Geistlinger *et al.* [2] found that self-contained methods, particularly GlobalTest and SAM-GS, were especially sensitive to gene set size, with a propensity toward detecting larger gene sets as significant (Supplementary Table 8 available online at <https://academic.oup.com/bib>). For example, even when random gene sets were assembled, GlobalTest and SAM-GS identified all gene sets with over 50 genes as significant. However, Maleki *et al.* [16] noted that GlobalTest was among the methods more likely to identify gene sets of smaller sizes as significant (Table 1; Supplementary Table 3 available online at <https://academic.oup.com/bib>), albeit, in this case, the upper bound for genes in a given gene set was nearly 2000, while in Geistlinger *et al.* [2], it was 500.

These contradictory findings are a prime example of the challenges associated with benchmarking methods for gene set analysis. Such glaring variability in results yielded by the same method investigated in different studies may be due to several factors, such as gene set size or differing proportions of DEGs in the studied datasets. For instance, GlobalTest tends to perform sub-optimally when only a few genes in a given gene set are differentially expressed and the majority of genes are not, and it conversely tends to be better suited for when there are many genes with small changes in differential expression in a gene set [37, 40]. We further discuss the impact of gene set size on results in a subsequent section as well as in Supplementary Text 1 (available online at <https://academic.oup.com/bib>).

If opting to select a competitive method instead, one must consider that testing the competitive null hypothesis often inherently implies the intended association not only between the phenotype and the genes within a given gene set but also between the phenotype and the genes in the complement of the set [40]. That said, competitive methods can be appropriate when the goal is to test for excessive amounts of differential expression among genes in a gene set. For instance, the popular ORA method was noted as suitable when there are large levels of differential expression [2]. However, ORA also tends to prioritize larger gene sets, assigning them lower *P*-values [16, 23]. Nonetheless, in Geistlinger *et al.* [2], ORA and other competitive methods outperformed the self-contained ones in ranking phenotype relevant gene sets near the top (Supplementary Table 9 available online at <https://academic.oup.com/bib>). In contrast, although ORA performed favorably on the prioritization of relevant gene sets in Tarca *et al.* [23], no clear discernment could be made with regard to the performance of competitive and self-contained methods on this measure (Supplementary Table 6 available online at <https://academic.oup.com/bib>). Furthermore, while self-contained methods tended to identify a larger proportion of gene sets as significant in Geistlinger *et al.* [2], the majority of competitive methods (i.e. SAFE [41],

GSEA, GSA and PADOG) did not identify any significant gene sets.

Intimately linked to the formulation of the null hypothesis is the calculation of the *P*-value [34]. Divergent approaches to assign a *P*-value to a gene set address the following question: What is the sampling unit? If the sampling unit is the gene, for each gene set of a given size, an equal number of genes are randomly drawn from all genes under investigation to sample the null distribution. If, however, the sampling unit is the subject, the phenotypic labels of subjects are randomly permuted to sample the null distribution instead. While methods that test a self-contained null hypothesis are generally linked with sample permutation and competitive methods with gene permutation, the latter group of methods can be modified to make them self-contained [40].

Sample permutation is often regarded as the preferred approach to obtain the empirical null distribution as its setup tends to pertain more naturally to the research question at hand of whether or not an association exists between a gene set and a phenotype. In contrast, methods that calculate significance by gene permutations suffer from the assumption that genes are independent and identically distributed (iid). It is well established, however, that this premise does not hold true in a real biological context where gene correlations (i.e. the coordinated expression of genes) can be observed and where sets of genes are known to work in tandem [37]. Thus, in the case of gene permutations, while significant gene sets may be reflective of either gene correlations that arise regardless of experimental condition and/or actual phenotypic differences, it is the latter that is often far more interesting, and the former can inflate the number of false positives [37, 40, 42, 43].

The effects of correlations within gene sets have been observed in various studies. Tamayo and colleagues [44] show that these correlations can have major implications on the results of enrichment analysis by comparing the results of GSEA against a simple parametric approach in 50 datasets. They observed that the parametric approach, which assumes differential gene expression scores are both independent and follow a normal distribution, yields a larger number of significant gene sets than GSEA, but many of these are speculated to be false positives. Similarly, in experiments on simulated data in Maciejewski [40], the author demonstrated that when gene correlations were present in the gene set yet there were no DEGs either in the gene set or its complement, false positive rates for methods that make the iid assumption (e.g. parametric methods proposed in Irizarry *et al.* [45] and competitive methods with gene permutation) were greater than expected. Thus, the authors of these studies caution that methods that assume gene independence may report gene sets as significantly associated with a phenotype when in fact gene correlations account for the purported, significant results. However, it is also worth noting that the influence of correlations can

be somewhat mitigated by reducing redundancies within gene sets.

In Maciejewski [40], the author observed that among methods with a sample permutation procedure, GlobalTest, GSEA and GSA and its variant achieved higher power. Furthermore, GSEA, a competitive method with sample permutation, had higher power than several other methods tested (i.e. GSA and its variant, PAGE [46], Wilcoxon rank sum test, Q1 [47] and SAFE), although as the number of DEGs in a gene set increased, so too did the power of the other methods.

Nevertheless, sample permutation requires an adequate number of samples as without a sufficiently large sample size, the calculated *P*-value may never achieve significance, in which case, gene permutation is recommended. For instance, in their comparative analysis, Maleki *et al.* [48] found that, across 10 replicate datasets, GSEA with sample permutation was unable to detect any gene set as enriched when sample sizes were small, suggesting a lower bound of 10 samples for this particular method. The robustness of various methods to changes in sample size is further discussed in [Supplementary Text 2](https://academic.oup.com/bib) (available online at <https://academic.oup.com/bib>).

Other methods have been proposed that attempt to address some of the drawbacks associated with sample and gene permutation approaches by conducting both sample permutations and gene randomizations in a method known as restandardization, as with GSA, through the use of rotations for gene set testing, as with FRY [49] and ROAST, or via bootstrapping methods, as in Zahn *et al.* [50] and Barry *et al.* [43].

Topology- and non-topology-based methods

Methods for enrichment analysis can also be classified as those which are topology-based or non-topology-based. The latter group of methods can be further sub-classified into the aforementioned ORA and FCS methods, the so-called first- and second-generation approaches, respectively [1]. PT- or topology-based methods fall into the category of third-generation approaches, intuitively more advanced as, unlike ORA and FCS methods, they leverage the topological structure of genes in a pathway. Nonetheless, results from multiple benchmarks on topology- and non-topology-based methods are inconclusive as to the superiority of one group of methods over another, with studies suggesting topology-based methods have the upper hand.

In Bayerlová *et al.* [13], authors noted that whether a method was topology-based or not was inconsequential to performance when original KEGG pathways (which tend to contain overlapping genes) were used in experiments ([Supplementary Tables 3–6](https://academic.oup.com/bib) available online at <https://academic.oup.com/bib>). Notably, while CePa includes pathways from both KEGG and the Pathway Interaction Database [51], other topology-based methods

evaluated in the study (i.e. PathNet and SPIA) are only compatible with pathways formatted in a custom-XML format (i.e. KEGG Markup Language). This result is particularly striking, considering KEGG contains overlapping pathways, thus limiting the potential of topology-based methods by restricting users to pathways formatted in the manner specified by this database. In contrast, experiments done using non-overlapping pathways resulted in topology-based methods outperforming non-topology-based ones [13]. In line with these findings, comparative studies by Jaakkola and Elo [14] and Nguyen *et al.* [9] similarly suggested that topology-based methods exhibit an improved performance over non-topology-based ones under certain conditions, albeit, contrary to findings by Bayerlová *et al.* [13], these conclusions were drawn exclusively using KEGG as the choice of pathway database.

More particularly, results from Nguyen *et al.* [9] indicate that topology-based methods have a slight upper hand in detecting target pathways as compared to non-topology-based ones ([Supplementary Table 6](https://academic.oup.com/bib) available online at <https://academic.oup.com/bib>), though results were mixed regarding the *P*-values of target pathways. In Jaakkola and Elo [14], topology-based methods (i.e. SPIA, CePa and NetGSA [52]) detected a larger number of significant pathways than non-topology-based ones (i.e. GSEA, Pathifier [53] and DAVID [54]). However, in a more challenging dataset where differences across groups were subtle, nearly all studied methods identified either no pathways or relatively few pathways as significantly enriched.

Ihnatova *et al.* [3] conducted several experiments, which assessed the influence of various parameters on topology-based methods [e.g. sensitivity to pathway and sample size ([Supplementary Table 7](https://academic.oup.com/bib) available online at <https://academic.oup.com/bib>), specificity ([Supplementary Table 5](https://academic.oup.com/bib) available online at <https://academic.oup.com/bib>) and exclusion of topological information]. As a proxy to study the latter parameter (i.e. whether topological information affects results for a given topological method), the authors evaluated the influence of single genes on the fraction of pathways that were considered enriched, assuming that a setup that fails to take into account PT is one in which individual genes have an equal impact on results. To that end, they found that TopologyGSA [55] and Clipper [56] yielded no difference in performance when topological information was excluded, while for all other methods, the exclusion of topological information led to the identification of a smaller fraction of enriched pathways. In addition, in assessing whether the ranks/*P*-values of target pathways change when topological information is incorporated, the authors found that both the ranks and *P*-values of target pathways decreased for PRS [57] and CePa, while for all other methods, the inclusion of topological information resulted in either no change or an increase in ranks/*P*-values of target pathways (at times caused by pathway-specific effects).

Additional methodological considerations and consensus approaches

Besides the above-mentioned common measures and classifications, several comparative studies have used to draw distinctions between enrichment methods, the performance of methods on a number of additional aspects has also been benchmarked. We refer to the studies that evaluate other aspects, including accuracy (Supplementary Table 10 available online at <https://academic.oup.com/bib>), type I error rate, power, runtime and assessments of reproducibility across datasets, among others in Supplementary Table 11 (available online at <https://academic.oup.com/bib>). Furthermore, we outline additional methodological considerations, including the steps used in data preprocessing and biases, which arise from experiments (Supplementary Text 3 available online at <https://academic.oup.com/bib>), the gene- and gene set-level statistics selected (Supplementary Texts 4 and 5 available online at <https://academic.oup.com/bib>), the applicability of enrichment analysis to various omics dataset types (Supplementary Text 6 available online at <https://academic.oup.com/bib>) and the choice of background (Supplementary Text 7 available online at <https://academic.oup.com/bib>).

Given the vast variety of enrichment methods, often with tunable settings, hundreds of methods and variants are at the disposal of life science researchers. As results can acutely vary according to the method selected, such a broad variability has prompted the development of tools to conduct enrichment analysis in concert. While the techniques to do so can differ, generally a consensus is taken across several methods to determine the final set of pathways that are interesting in some statistically significant way. Examples to do so include the R packages EGSEA [58], EnrichmentBrowser [59], Piano [60] and decoupleR [61] as well as the ML-based approach, CGPS [62] and the CPA web application [63]. Details on each of these ensemble techniques are provided in Supplementary Text 8 (available online at <https://academic.oup.com/bib>).

Impact of pathway database and gene set size

While variations of enrichment methods have been among the most studied factors that influence the results of an enrichment analysis, there are several other considerations to be made in the design of an experiment to ensure biologically meaningful results. In this section, we introduce studies, including notable benchmarks, that have investigated the impact of additional factors on the results of enrichment analysis, such as database choice and pathway size.

One of the most critical factors the results of an enrichment analysis can hinge upon is the choice of a reference pathway database(s). It is common practice for researchers to solely rely upon a single database for an enrichment analysis, which can be due, in part, to

a researcher's preferences, the popularity of a particular database or its ease of usage, among other factors. Indeed, we observed that the majority of studies that benchmarked the performance of enrichment methods (Table 1) were almost always conducted on a single database, and that too, primarily KEGG.

A first investigation on the importance of selecting a collection of gene sets was performed by Bateman *et al.* [64]. In this study, the authors demonstrated how the seven standard collections housed within MSigDB yielded different results when conducting GSEA within the context of a drug response cancer dataset. Among other findings, the results of this study indicated that some collections were able to yield a significantly larger number of enriched pathways relevant to the studied phenotype than others. Furthermore, the authors argued that the choice of gene set collections should not be made arbitrarily as certain gene sets may be more or less suitable for a particular dataset than others. In a recent study on best practices for the popular ORA method on metabolomics data [65], the authors also found that the results of pathway analysis substantially differed based on the choice of pathway database (i.e. KEGG, Reactome and BioCyc [66]).

Similar conclusions were drawn in our previous work [67] in which we evaluated whether enrichment results are in consensus for any given pathway that can be found across three major pathway databases (i.e. KEGG, Reactome and WikiPathways [68]) and multiple enrichment methods. Our study revealed the advantages of combining multiple databases by using equivalent pathway mappings, demonstrating that an integrative resource can yield more consistent results than an individual one. Overall, these studies demonstrate the importance of database choice, a crucial factor given the differences in coverage across databases [69, 70]. Finally, we would also like to note the importance of database size as the total number of pathways present in a database has an influence when multiple correction methods are applied.

An additional factor that is related to database choice is gene set (pathway) size, corresponding to the number of genes within a gene set for enrichment methods that do not consider PT, or the number of nodes (genes) and edges for those that do consider it. The effect of pathway size has recently been studied in Karp *et al.* [71] by comparing the significance of six equivalent pathway definitions from KEGG and EcoCyc [72]. Given the differences in the average size of a pathway across the two databases (i.e. KEGG pathways are significantly larger than their respective homologs in EcoCyc), the authors investigated the degree to which size could influence results, finding that pathway size can have a stronger effect than the statistical corrections used. Furthermore, the authors found that KEGG pathways required up to two times as many significant genes in order to attain the same *P*-value as their EcoCyc counterparts.

Notably, size differences between equivalent pathways have not only been examined for these two databases but

also across other major resources, such as Reactome, and WikiPathways. In this work, the authors argue that using pathway definitions that span across several biological processes (e.g. signal transduction) can lead to misinterpretations as when these pathways are enriched, it is difficult to construe whether this implicates all or only a subset of the pathway. These broadly defined pathways can also be less informative, contributing little in terms of novelty to the overall understanding of the distinctions between the phenotypes under study. Nonetheless, smaller pathways can lead to exceedingly long results and overly strict multiple testing corrections [4].

Possible solutions for mitigating the impact of gene set size on results are defining the minimum and maximum number of genes within a gene set (e.g. between 10 and 500), careful consideration of the enrichment analysis method selected (see 'Hypothesis testing and significance assessment' section) as well as addressing redundancies within gene sets, as proposed in [73]. In their approach, the authors suggest discarding significant gene sets that overlap with others in order to ensure that the enrichment of a particular pathway is not a result of the overlay.

While database choice and pathway size are two critical factors to consider, we foresee several approaches to offset the challenges they create. In the case of database choice, a study by Maleki *et al.* [74] proposed two simple metrics (i.e. permeability and maximum achievable coverage scores) to assess the degree of overlap between a gene list of relevance and all gene sets within a database. The goal of these metrics is to provide an intuition of whether or not the genes of a phenotype under investigation are well covered by a particular database. Thus, the authors argue that this approach can reduce database bias and arbitrary database selection as the two scores can guide users to rationally decide upon the most appropriate database.

Another solution that we propose is that the enrichment results generated from a reference database could be validated against an additional database using equivalent pathway mappings across them. By leveraging pathway mappings, one can assess the similarity between the results obtained from different databases (i.e. reference and 'validation' database) to confirm whether they are in consensus, or re-evaluate them if they are not. In earlier work, we leveraged this technique by generating equivalent pathway mappings across four pathway databases [75]. A web tool (i.e. DecoPath) subsequently enables users to evaluate similarities and differences at the gene and pathway level for a given pathway across databases and enrichment methods. For instance, a particular pathway in one database can have a slightly different gene set than the same pathway in another database, which can ultimately explain why a pathway is detected as significantly enriched in one database but not in another.

Similarly, pathway mappings can also be employed to systematically study the impact of pathway size on results. Here, one could leverage hierarchical mappings

(i.e. pathway A is part of pathway B) from pathway ontologies to evaluate whether related pathways are similarly enriched. Although a pathway ontology was earlier proposed by [76], it has neither been adopted by nor linked to any major database. Instead, each database utilizes its own pathway terminology, though some databases such as Reactome and GO also incorporate a hierarchical organization within their schema. In fact, Reactome recently adopted such an approach to facilitate the interpretation of enrichment analyses through implementing ReacFoam, a visualization for navigating through its pathway hierarchy and exploring the degree of enrichment of pathways at different levels.

The growth of biomedical literature is reflected in pathway databases as their pathway definitions change over time. A study by Wadi *et al.* [77] demonstrated the impact of outdated pathway definitions in several web-based tools as well as highlighted that the number of pathways/biological processes doubled in 7 years (2009–16) in major resources such as Reactome and GO. Furthermore, it revealed that the majority of the studies analyzed were conducted using outdated pathway definitions, constituting a major issue as the results presented in such studies could have potentially changed. We believe this problem can be partially mitigated if users are alerted by pathway enrichment tools when the underlying pathway database(s) has not been recently updated. Furthermore, updating the information from pathway databases in a tool has been greatly simplified by the APIs and services offered by major resources such as Reactome, GO and WikiPathways. Finally, we encourage researchers to include both the version of the database(s) used in the analysis as well as the version of the tool(s) employed.

Impact of additional factors on enrichment analysis and possible future benchmarks

While the factors mentioned thus far have each been benchmarked with regard to their impact on pathway enrichment results, there exist other factors that have not yet been explored in detail. First, at a more granular level, individual genes can also have an impact on results. A study by Ballouz *et al.* [78] raised the challenges associated with annotation bias and redundancies in gene sets. The annotation of a single gene to many functions (i.e. multifunctional genes) can potentially confound the results of a pathway analysis as these genes may result in a sizeable number of enriched pathways that are largely irrelevant. For example, several pathways with multifunctional genes may be considered enriched in the results, though the enrichment of these pathways could be due to the presence of multifunctional genes rather than the relevance of the pathway to the phenotype of interest. One approach the authors propose to control this effect is by performing repeated runs of the analysis while removing the topmost multifunctional genes in the dataset in order to identify the most robust pathways.

Furthermore, other ways to reduce the effect of multi-functional genes can include assigning weights to genes based on their promiscuity, though this approach might also have drawbacks.

A second factor that has not yet been investigated, which is related both to database updates and choice, is the size of a database measured by the number of pathways. This factor is not only important due to its correlation with the coverage of biological processes but also because the size of the database can influence the significance of the results when correcting for multiple testing (see [Supplementary Text 9](#) available online at <https://academic.oup.com/bib>). As a consequence, depending on the size of a database, the same pathway in one database may or may not be enriched in another after applying multiple testing correction. This is often the case when comparing popular databases, such as KEGG and Reactome, whose number of pathways can differ by an order of magnitude.

Finally, we would like to note that there are other interesting factors, which could potentially be analyzed in the future. First, for topology-based methods, the particular network structure of some pathways may make them more susceptible to enrichment than others given the topological differences identified by [79]. Thus, one future possible benchmark could investigate the effect of network sparsity on pathway enrichment, or if hubs within a network correlate with greater enrichment. Second, another factor to evaluate is the degree to which a bias toward certain indications in pathway knowledge influences results. For example, there is an over-representation of interactions characterized in widely studied indication areas, such as cancer [80, 81], and thus, pathways containing these interactions may appear in the results of enrichment, while possessing little relevance to the studied phenotype. To investigate this factor, resources such as BioGrid [82] where protein–protein interactions are annotated with experimental metadata can be leveraged since the majority of databases do not provide information on the provenance supporting each interaction. Third, only a minute fraction of known proteins have been experimentally annotated with functional characterizations, while functional annotations for the vast majority of proteins are either inferred, presumptive or unknown [83, 84]. Several computational methods exist for protein function prediction, and while such methods are routinely benchmarked [85], the effect of experimental versus predicted functional annotations of proteins on downstream analyses also warrants further study. This is of particular importance to GO enrichment, where numerous algorithms have been developed to predict GO terms for proteins [86].

Discussion

The last decade has seen an explosion in the usage of pathway enrichment analysis, spearheaded by both an abundance in the volume of available data and the

interpretive power of these analyses [10]. Prompted by a wide range of available enrichment methods and pathway resources, several comparative studies have evaluated how different factors can influence the results of such an analysis. Here, we have reviewed the findings of these studies in order to provide a comprehensive overview on the impact of these factors. Furthermore, we have suggested possible approaches to overcome some of the limitations discussed as well as possibilities for additional benchmark studies on other, under studied factors.

In the first section of this review, we have outlined the results of 12 comparative studies that have investigated differences across pathway enrichment methods. Many of these studies have specifically focused on the performance of individual methods on popular metrics (e.g. prioritization, sensitivity and specificity), keeping in mind that without gold standards to conclude whether the results from any given method are biologically sound, objective evaluations can be difficult to make. Overall, we have found many inconsistencies in the performance of methods across metrics as well as across studies. While there is no consensus across studies on whether a specific method outperforms others, we have reported trends we have observed regarding the top-performing methods for each metric.

Though we note that the performance of the majority of methods on these and other metrics is inconclusive, whether a particular method is a reasonable choice for a certain use case can depend on a number of factors, such as the goal of the experiment, the dataset in question or particulars of the gene set collection. Nevertheless, trade-offs between performances on certain metrics can be important considerations in the selection of a method. For example, given a dataset where changes in differential gene expression between experimental groups are subtle, a highly sensitive method can increase the likelihood of detecting a signal. Thus, a large number of gene sets that are truly significant can be identified, essentially ruling out nearly all gene sets that are not detected, albeit at the expense of producing a greater number of false positives. If, however, changes in differential gene expression between experimental groups are generally more pronounced, a method ranked high in specificity may be preferable to preclude the detection of too many gene sets, which can complicate interpretation.

We have also examined comparative studies that have evaluated the differences between distinct categories of enrichment methods, such as how the null hypothesis is formulated and the sampling unit is defined, noting that the selection of one category of methods over another can have serious repercussions on the fraction of gene sets that are significant and their ranks. In addition, a major categorical distinction is drawn between topology- and non-topology-based methods, which have been reviewed in several benchmarks. We have found that, though topology-based approaches are more advanced, for some methods, the removal

of topological information yields no differences in results, for other methods, it can improve results, and several are constrained in that they only cater to KEGG pathways (or pathways in an equivalent format). Finally, we reviewed studies that have assessed the influence of particular, modular aspects of a typical enrichment analysis as well as outlined additional aspects one must be cognizant of that can affect the behavior of a given method, which ultimately reflects in the overall results of an analysis.

We have reviewed several other factors apart from enrichment methods, such as pathway size and database choice. Notably, the latter can be subjective, with both researcher preferences and distinct research goals taking precedence over set guidelines. However, we have outlined approaches that leverage pathway mappings to mitigate the effect of these factors. An additional aspect discussed in this review is the lack of regular updates to enrichment tools, which reflect updates made to pathway databases. Fortunately, this issue has, at least, partially been addressed by the adoption of API services by major pathway resources. Nevertheless, the amount of literature published on a daily basis continues to grow, making the task of maintaining up-to-date pathway definitions difficult, particularly for public and academic resources. Thus, we envisage that the path forward to address this shortfall is to improve interoperability across databases via mappings [70] or through the use of common database formats [87].

Finally, we would like to mention possible future benchmarks beyond the ones we have previously proposed. First, future benchmarks can benefit from the existence of a gold standard prioritization approach, for instance, one that leverages well-established pathway-disease associations from genetic disorders, similar to the assessment proposed in [9], which exploits knockout datasets. Second, given the rise of multi-omics datasets, we anticipate the development of enrichment methods that operate on other modalities beyond mRNA data, such as metabolomics (see [Supplementary Text 6](#) available online at <https://academic.oup.com/bib>). Last, we foresee that the insights gained from multi-omics experiments will also be reflected in pathway definitions in two ways: (i) the appearance of ‘dynamic pathways’ (i.e. contextualized pathways representing particular pathway states as opposed to general, static diagrams) and (ii) a shift from traditional gene sets to sets of multimodal biological entities.

Conclusion

In conclusion, the effect of various factors on pathway enrichment analysis is apparent. Numerous studies have demonstrated how variations in the design of an enrichment analysis can lead to altogether different findings. At the extremes, comparative studies have shown how certain experimental setups can detect either all or no

gene sets as interesting in some statistically significant way. We summarize the key findings of studies reviewed herein as follows:

Formulation of null hypothesis and significance assessment

One must be cognizant of how the null hypothesis is formulated (i.e. competitive or self-contained) as methods categorized into one or another approach behave differently in terms of the fraction of gene sets reported as significant, as well as their sensitivity to gene set size, sample size and sample heterogeneity. Self-contained methods also tend to have greater power than competitive methods and careful consideration should be made taking into account the proportion of genes that are differentially expressed in the dataset. Similarly, in order to calculate a *P*-value for each gene set, one must bear in mind that disparate approaches can impact the results of an enrichment analysis, and depending on the approach taken, introduce false positives.

Pathway and sample size considerations

Certain enrichment methods have been observed to be more or less robust to pathway and sample size than certain others. Sensitive methods may detect larger gene sets as significantly enriched and their sensitivity can be tied with whether they are competitive or self-contained methods. Not surprisingly, a method’s performance tends to deteriorate with decreasing sample size, although some methods are more robust on this factor than others.

Topology- versus non-topology-based methods

Topology-based methods are intuitively more advanced than non-topology-based ones. Incorporation of topological information tends to improve the ranks and *P*-values of relevant pathways for some topology-based methods, yet this may not be the case for all. Nonetheless, some topology-based methods are limited or at least partial to specific pathway databases.

Choice of gene set collection or pathway database

The selection of one gene set collection over another can lead to different results. Some collections or databases may be more suitable than others for a given dataset. The selection of a database is complicated by variable definitions of pathway boundaries as well as by redundancies and outdated pathway definitions.

The errors from these steps that propagate through an enrichment analysis may be inconsequential at best and misleading at worst. Although there is no singular method or gene set collection/pathway database, which is advisable for enrichment analysis over all others, well-informed choices can be made and solutions to mitigate the impact of various factors are available. Furthermore, recently, many ensemble approaches have been developed so that users can benefit from multiple databases and/or methods.

Key Points

- Pathway enrichment analysis is a widely used technique for the interpretation of biological data
- In recent years, the advent of a multitude of enrichment methods and pathway databases has led to several benchmarks to study the impact of various factors on the results of enrichment analysis
- This review outlines key aspects of enrichment analysis and summarizes results of studies, which have evaluated their influence
- We propose solutions to mitigate the effect of these factors and identify possible future benchmarks

Authors' contributions

S.M. and D.D.-F. wrote the manuscript. A.T.K. and M.H.-A. reviewed the manuscript. All authors have read and approved the final manuscript.

Supplementary Data

Supplementary data are available online at [https://academic.oup.com/bib](https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bb1113).

Funding

The Fraunhofer Cluster of Excellence 'Cognitive Internet Technologies'.

References

1. Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput Biol* 2012;**8**(2):e1002375. <https://doi.org/10.1371/journal.pcbi.1002375>.
2. Geistlinger L, Csaba G, Santarelli M, et al. Toward a gold standard for benchmarking gene set enrichment analysis. *Brief Bioinform* 2020;**22**(1):545–56. <https://doi.org/10.1093/bib/bbz158>.
3. Ihnatova I, Popovici V, Budinska E. A critical comparison of topology-based pathway analysis methods. *PLoS One* 2018;**13**(1):e0191154. <https://doi.org/10.1371/journal.pone.0191154>.
4. Reimand J, Isserlin R, Voisin V, et al. Pathway enrichment analysis and visualization of omics data using g: Profiler, GSEA, Cytoscape and EnrichmentMap. *Nat Protoc* 2019;**14**(2):482–517. <https://doi.org/10.1038/s41596-018-0103-9>.
5. Kanehisa M, Furumichi M, Sato Y, et al. KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res* 2021;**49**(D1):D545–51. <https://doi.org/10.1093/nar/gkaa970>.
6. Fabregat A, Jupe S, Matthews L, et al. The reactome pathway knowledgebase. *Nucleic Acids Res* 2018;**46**(D1):D649–55. <https://doi.org/10.1093/nar/gkx1132>.
7. The Gene Ontology Consortium. The Gene Ontology resource: enriching a GOLD mine. *Nucleic Acids Res* 2021;**49**(D1):D325–34. <https://doi.org/10.1093/nar/gkaa1113>.
8. Maleki F, Ovens K, Hogan DJ, et al. Gene set analysis: challenges, opportunities, and future research. *Front Genet* 2020;**11**:654. <https://doi.org/10.3389/fgene.2020.00654>.
9. Nguyen TM, Shafi A, Nguyen T, et al. Identifying significantly impacted pathways: a comprehensive review and assessment. *Genome Biol* 2019;**20**(1):1–15. <https://doi.org/10.1186/s13059-019-1790-4>.
10. Xie C, Jauhari S, Mora A. Popularity and performance of bioinformatics software: the case of gene set analysis. *BMC Bioinform* 2021;**22**(1):1–16. <https://doi.org/10.1186/s12859-021-04124-5>.
11. Massey FJ, Jr. The Kolmogorov-Smirnov test for goodness of fit. *J Am Stat Assoc* 1951;**46**(253):68–78. <https://doi.org/10.1080/01621459.1951.10500769>.
12. Wilcoxon F. Individual comparisons by ranking methods. In *Breakthroughs in Statistics*. New York: Springer, 1992, 196–202. <https://doi.org/10.2307/3001968>.
13. Bayerlová M, Jung K, Kramer F, et al. Comparative study on gene set and pathway topology-based enrichment methods. *BMC Bioinform* 2015;**16**(1):334. <https://doi.org/10.1186/s12859-015-0751-5>.
14. Jaakkola MK, Elo LL. Empirical comparison of structure-based pathway methods. *Brief Bioinform* 2016;**17**(2):336–45. <https://doi.org/10.1093/bib/bbv049>.
15. Ma J, Shojaie A, Michailidis G. A comparative study of topology-based pathway enrichment analysis methods. *BMC Bioinform* 2019;**20**(1):1–14. <https://doi.org/10.1186/s12859-019-3146-1>.
16. Maleki F, Ovens KL, Hogan DJ, et al. Measuring consistency among gene set analysis methods: a systematic study. *J Bioinform Comput Biol* 2019a;**17**(05):1940010. <https://doi.org/10.1142/S0219720019400109>.
17. Liberzon A, Subramanian A, Pinchback R, et al. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 2011;**27**(12):1739–40. <https://doi.org/10.1093/bioinformatics/btr260>.
18. Mathur R, Rotroff D, Ma J, et al. Gene set analysis methods: a systematic comparison. *BioData Mining* 2018;**11**(1):1–19. <https://doi.org/10.1186/s13040-018-0166-8>.
19. Dutta B, Wallqvist A, Reifman J. PathNet: a tool for pathway analysis using topological information. *Source Code Biol Med* 2012;**7**(1):1–12. <https://doi.org/10.1186/1751-0473-7-10>.
20. Gu Z, Liu J, Cao K, et al. Centrality-based pathway enrichment: a systematic approach for finding significant pathways dominated by key genes. *BMC Syst Biol* 2012;**6**(1):1–13. <https://doi.org/10.1186/1752-0509-6-56>.
21. Gu Z, Wang J. CePa: an R package for finding significant pathways weighted by multiple network centralities. *Bioinformatics* 2013;**29**(5):658–60. <https://doi.org/10.1093/bioinformatics/btt008>.
22. Rahmatallah Y, Emmert-Streib F, Glazko G. Gene set analysis approaches for RNA-seq data: performance evaluation and application guideline. *Brief Bioinform* 2016;**17**(3):393–407. <https://doi.org/10.1093/bib/bbv069>.
23. Tarca AL, Bhatti G, Romero R. A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity. *PLoS One* 2013;**8**(11):e79217. <https://doi.org/10.1371/journal.pone.0079217>.
24. Zyla J, Marczyk M, Polanska J. Reproducibility of finding enriched gene sets in biological data analysis. In: *International Conference on Practical Applications of Computational Biology & Bioinformatics*, Porto, Portugal. Cham: Springer International Publishing, 2017. pp. 146–54. https://doi.org/10.1007/978-3-319-60816-7_18.
25. Zyla J, Marczyk M, Domaszewska T, et al. Gene set enrichment for reproducible science: comparison of CERNO and

- eight other algorithms. *Bioinformatics* 2019;**35**(24):5146–54. <https://doi.org/10.1093/bioinformatics/btz447>.
26. Michaud J, Simpson KM, Escher R, et al. Integrative analysis of RUNX1 downstream pathways and target genes. *BMC Genomics* 2008;**9**(1):1–17. <https://doi.org/10.1186/1471-2164-9-363>.
 27. Goeman JJ, Van De Geer SA, De Kort F, et al. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* 2004;**20**(1):93–9. <https://doi.org/10.1093/bioinformatics/btg382>.
 28. Tomfohr J, Lu J, Kepler TB. Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinform* 2005;**6**(1):1–11. <https://doi.org/10.1186/1471-2105-6-225>.
 29. Tarca AL, Draghici S, Khatri P, et al. A novel signaling pathway impact analysis. *Bioinformatics* 2009;**25**(1):75–82. <https://doi.org/10.1093/bioinformatics/btn577>.
 30. Wu D, Smyth GK. Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Res* 2012;**40**(17):e133–3. <https://doi.org/10.1093/nar/gks461>.
 31. Efron B, Tibshirani R. On testing the significance of sets of genes. *Ann Appl Stat* 2007;**1**(1):107–29. <https://doi.org/10.1214/07-AOAS101>.
 32. Tarca AL, Draghici S, Bhatti G, et al. Down-weighting overlapping genes improves gene set analysis. *BMC Bioinform* 2012;**13**(1):1–14. <https://doi.org/10.1186/1471-2105-13-136>.
 33. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci* 2005;**102**(43):15545–50. <https://doi.org/10.1073/pnas.0506580102>.
 34. Goeman JJ, Bühlmann P. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics* 2007;**23**(8):980–7. <https://doi.org/10.1093/bioinformatics/btm051>.
 35. Ackermann M, Strimmer K. A general modular framework for gene set enrichment analysis. *BMC Bioinform* 2009;**10**(1):1–20. <https://doi.org/10.1186/1471-2105-10-47>.
 36. Tripathi S, Glazko GV, Emmert-Streib F. Ensuring the statistical soundness of competitive gene set approaches: gene filtering and genome-scale coverage are essential. *Nucleic Acids Res* 2013;**41**(7):e82–2. <https://doi.org/10.1093/nar/gkt054>.
 37. Wu MC, Lin X. Prior biological knowledge-based approaches for the analysis of genome-wide expression profiles using gene sets and pathways. *Stat Methods Med Res* 2009;**18**(6):577–93. <https://doi.org/10.1177/0962280209351925>.
 38. Wu D, Lim E, Vaillant F, et al. ROAST: rotation gene set tests for complex microarray experiments. *Bioinformatics* 2010;**26**(17):2176–82. <https://doi.org/10.1093/bioinformatics/btq401>.
 39. Dinu I, Potter JD, Mueller T, et al. Improving gene set analysis of microarray data by SAM-GS. *BMC Bioinform* 2007;**8**(1):1–13. <https://doi.org/10.1186/1471-2105-8-242>.
 40. Maciejewski H. Gene set analysis methods: statistical models and methodological differences. *Brief Bioinform* 2014;**15**(4):504–18. <https://doi.org/10.1093/bib/bbt002>.
 41. Barry WT, Nobel AB, Wright FA. Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics* 2005;**21**(9):1943–9. <https://doi.org/10.1093/bioinformatics/bti260>.
 42. Nam D. Effect of the absolute statistic on gene-sampling gene-set analysis methods. *Stat Methods Med Res* 2017;**26**(3):1248–60. <https://doi.org/10.1177/0962280215574014>.
 43. Barry WT, Nobel AB, Wright FA. A statistical framework for testing functional categories in microarray data. *Ann Appl Stat* 2008;**2**(1):286–315. <https://doi.org/10.1214/07-AOAS146>.
 44. Tamayo P, Steinhardt G, Liberzon A, et al. The limitations of simple gene set enrichment analysis assuming gene independence. *Stat Methods Med Res* 2016;**25**(1):472–87. <https://doi.org/10.1177/0962280212460441>.
 45. Irizarry RA, Wang C, Zhou Y, et al. Gene set enrichment analysis made simple. *Stat Methods Med Res* 2009;**18**(6):565–75. <https://doi.org/10.1177/0962280209351908>.
 46. Kim SY, Volsky DJ. PAGE: parametric analysis of gene set enrichment. *BMC Bioinform* 2005;**6**(1):1–12. <https://doi.org/10.1186/1471-2105-6-144>.
 47. Tian L, Greenberg SA, Kong SW, et al. Discovering statistically significant pathways in expression profiling studies. *Proc Natl Acad Sci* 2005;**102**(38):13544–9. <https://doi.org/10.1073/pnas.0506577102>.
 48. Maleki F, Ovens K, McQuillan I, et al. Sample size and reproducibility of gene set analysis. In: *IEEE International Conference on Bioinformatics and Biomedicine, Madrid, Spain*. New York, NY, USA: IEEE, 2018. pp. 122–9. <https://doi.org/10.1109/BIBM.2018.8621462>.
 49. Ritchie ME, Phipson B, Wu DI, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015;**43**(7):e47–7. <https://doi.org/10.1093/nar/gkv007>.
 50. Zahn JM, Sonu R, Vogel H, et al. Transcriptional profiling of aging in human muscle reveals a common aging signature. *PLoS Genet* 2006;**2**(7):e115. <https://doi.org/10.1371/journal.pgen.0020115>.
 51. Schaefer CF, Anthony K, Krupa S, et al. PID: the pathway interaction database. *Nucleic Acids Res* 2009;**37**(suppl_1):D674–9. <https://doi.org/10.1093/nar/gkn653>.
 52. Shojaie A, Michailidis G. Network enrichment analysis in complex experiments. *Stat Appl Genet Mol Biol* 2010;**9**(1). <https://doi.org/10.2202/1544-6115.1483>.
 53. Drier Y, Sheffer M, Domany E. Pathway-based personalized analysis of cancer. *Proc Natl Acad Sci* 2013;**110**(16):6388–93. <https://doi.org/10.1073/pnas.1219651110>.
 54. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 2009;**4**(1):44–57. <https://doi.org/10.1038/nprot.2008.211>.
 55. Massa MS, Chiogna M, Romualdi C. Gene set analysis exploiting the topology of a pathway. *BMC Syst Biol* 2010;**4**(1):1–15. <https://doi.org/10.1186/1752-0509-4-121>.
 56. Martini P, Sales G, Massa MS, et al. Along signal paths: an empirical gene set approach exploiting pathway topology. *Nucleic Acids Res* 2013;**41**(1):e19–9. <https://doi.org/10.1093/nar/gks866>.
 57. Ibrahim MAH, Jassim S, Cawthorne MA, et al. A topology-based score for pathway enrichment. *J Comput Biol* 2012;**19**(5):563–73. <https://doi.org/10.1089/cmb.2011.0182>.
 58. Alhamdoosh M, Ng M, Wilson NJ, et al. Combining multiple tools outperforms individual methods in gene set enrichment analyses. *Bioinformatics* 2017;**33**(3):414–24. <https://doi.org/10.1093/bioinformatics/btw623>.
 59. Geistlinger L, Csaba G, Zimmer R. Bioconductor's Enrichment-Browser: seamless navigation through combined results of set- & network-based enrichment analysis. *BMC Bioinform* 2016;**17**(1):1–11. <https://doi.org/10.1186/s12859-016-0884-1>.
 60. Våremo L, Nielsen J, Nookaew I. Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. *Nucleic Acids Res* 2013;**41**(8):4378–91. <https://doi.org/10.1093/nar/gkt111>.
 61. Badia-i-Mompel P, Vélez J, Braunger J, et al. decoupleR: ensemble of computational methods to infer biological activities from omics data. *Bioinformatics Advances*. 2022;**2**(1):vbac016. <https://doi.org/10.1093/bioadv/vbac016>.

62. Ai C, Kong L. CGPS: a machine learning-based approach integrating multiple gene set analysis tools for better prioritization of biologically relevant pathways. *J Genet Genomics* 2018;**45**(9):489–504. <https://doi.org/10.1016/j.jgg.2018.08.002>.
63. Nguyen H, Tran D, Galazka JM, et al. CPA: a web-based platform for consensus pathway analysis and interactive visualization. *Nucleic Acids Res* 2021;**49**(W1):W114–W124. <https://doi.org/10.1093/nar/gkab421>.
64. Bateman AR, El-Hachem N, Beck AH, et al. Importance of collection in gene set enrichment analysis of drug response in cancer cell lines. *Sci Rep* 2014;**4**:4092. <https://doi.org/10.1038/srep04092>.
65. Wieder C, Frainay C, Poupin N, et al. Pathway analysis in metabolomics: recommendations for the use of over-representation analysis. *PLoS Comput Biol* 2021;**17**(9):e1009105. <https://doi.org/10.1371/journal.pcbi.1009105>.
66. Karp PD, Billington R, Caspi R, et al. The BioCyc collection of microbial genomes and metabolic pathways. *Brief Bioinform* 2019;**20**(4):1085–93. <https://doi.org/10.1093/bib/bbx085>.
67. Mubeen S, Hoyt CT, Gemünd A, et al. The impact of pathway database choice on statistical enrichment analysis and predictive modeling. *Front Genet* 2019;**10**:1203. <https://doi.org/10.3389/fgene.2019.01203>.
68. Martens M, Ammar A, Riutta A, et al. WikiPathways: connecting communities. *Nucleic Acids Res* 2021;**49**(D1):D613–21. <https://doi.org/10.1093/nar/gkaa1024>.
69. Stobbe MD, Houten SM, Jansen GA, et al. Critical assessment of human metabolic pathway databases: a stepping stone for future integration. *BMC Syst Biol* 2011;**5**(1):165. <https://doi.org/10.1186/1752-0509-5-165>.
70. Domingo-Fernández D, Hoyt CT, Bobis-Álvarez C, et al. Com-Path: an ecosystem for exploring, analyzing, and curating mappings across pathway databases. *NPJ Syst Biol Appl* 2018;**4**(1):43. <https://doi.org/10.1038/s41540-018-0078-8>.
71. Karp PD, Midford PE, Caspi R, et al. Pathway size matters: the influence of pathway granularity on over-representation (enrichment analysis) statistics. *BMC Genomics* 2021;**22**(1):1–11. <https://doi.org/10.1186/s12864-021-07502-8>.
72. Keseler IM, Mackie A, Santos-Zavaleta A, et al. The EcoCyc database: reflecting new knowledge about *Escherichia coli* K-12. *Nucleic Acids Res* 2017;**45**(D1):D543–50. <https://doi.org/10.1093/nar/gkw1003>.
73. Simillion C, Liechti R, Lischer HE, et al. Avoiding the pitfalls of gene set enrichment analysis with SetRank. *BMC Bioinform* 2017;**18**(1):1–14. <https://doi.org/10.1186/s12859-017-1571-6>.
74. Maleki F, Ovens K, McQuillan I, et al. Gene set databases: A fountain of knowledge or a siren call? In: *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, Niagara Falls, NY, USA. New York, NY, USA: Association for Computing Machinery, 2019. pp. 269–78. <https://doi.org/10.1145/3307339.3342146>.
75. Mubeen S, Bharadhwaj VS, Gadiya Y, et al. DecoPath: a web application for decoding pathway enrichment analysis. *NAR Genomics Bioinform* 2021;**3**(3):lqab087. <https://doi.org/10.1093/nargab/lqab087>.
76. Petri V, Jayaraman P, Tutaj M, et al. The pathway ontology–updates and applications. *J Biomed Semant* 2014;**5**(1):1–12. <https://doi.org/10.1186/2041-1480-5-7>.
77. Wadi L, Meyer M, Weiser J, et al. Impact of outdated gene annotations on pathway enrichment analysis. *Nat Methods* 2016;**13**(9):705. <https://doi.org/10.1038/nmeth.3963>.
78. Ballouz S, Pavlidis P, Gillis J. Using predictive specificity to determine when gene set analysis is biologically meaningful. *Nucleic Acids Res* 2017;**45**(4):e20–0. <https://doi.org/10.1093/nar/gkw957>.
79. Rubel T, Singh P, Ritz A. Reconciling signaling pathway databases with network topologies. *Pac Symp Biocomput*. 2021;**27**:211–22. https://doi.org/10.1142/9789811250477_0020.
80. Reyes-Aldasoro CC. The proportion of cancer-related entries in PubMed has increased considerably; is cancer truly “The Emperor of All Maladies”? *PLoS One* 2017;**12**(3):e0173671. <https://doi.org/10.1371/journal.pone.0173671>.
81. Hanspers K, Riutta A, Summer-Kutmon M, et al. Pathway information extracted from 25 years of pathway figures. *Genome Biol* 2020;**21**(1):1–18. <https://doi.org/10.1186/s13059-020-02181-2>.
82. Oughtred R, Stark C, Breitkreutz BJ, et al. The BioGRID interaction database: 2019 update. *Nucleic Acids Res* 2019;**47**(D1):D529–41. <https://doi.org/10.1093/nar/gkw1102>.
83. Erdin S, Lisewski AM, Lichtarge O. Protein function prediction: towards integration of similarity metrics. *Curr Opin Struct Biol* 2011;**21**(2):180–8. <https://doi.org/10.1016/j.sbi.2011.02.001>.
84. Shehu A, Barbará D, Molloy K. A survey of computational methods for protein function prediction. In: *Big Data Analytics in Genomics*. Cham: Springer, 2016, 225–98 https://doi.org/10.1007/978-3-319-41279-5_7.
85. Zhou N, Jiang Y, Bergquist TR, et al. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biol* 2019;**20**(1):1–23. <https://doi.org/10.1186/s13059-019-1835-8>.
86. Makrodimitis S, van Ham RC, Reinders MJ. Improving protein function prediction using protein sequence and GO-term similarities. *Bioinformatics* 2019;**35**(7):1116–24. <https://doi.org/10.1093/bioinformatics/bty751>.
87. Good BM, Van Auken K, Hill DP, et al. Reactome and the Gene Ontology: digital convergence of data resources. *Bioinformatics* 2021;**37**(19):3343–48. <https://doi.org/10.1093/bioinformatics/btab325>.

A.2 The Impact of Pathway Database Choice on Statistical Enrichment Analysis and Predictive Modeling

Reprinted with permission from “Mubeen S., Hoyt C. T., Gemünd A., Hofmann-Apitius M., Fröhlich H., and Domingo-Fernández D. (2019). The Impact of Pathway Database Choice on Statistical Enrichment Analysis and Predictive Modeling. *Frontiers in Genetics*, 10:1203”.

Copyright © Mubeen, S., *et al.*, 2019



The Impact of Pathway Database Choice on Statistical Enrichment Analysis and Predictive Modeling

Sarah Mubeen^{1,2}, Charles Tapley Hoyt^{1,2†}, André Gemünd¹, Martin Hofmann-Apitius^{1,2}, Holger Fröhlich² and Daniel Domingo-Fernández^{1,2*}

¹ Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Sankt Augustin, Germany, ² Bonn-Aachen International Center for IT, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn, Germany

OPEN ACCESS

Edited by:

Lavanya Balakrishnan,
Mazumdar Shaw Medical Centre,
India

Reviewed by:

George C. Tseng,
University of Pittsburgh,
United States
Inyoung Kim,
Virginia Tech,
United States

*Correspondence:

Daniel Domingo-Fernández
daniel.domingo.fernandez@scai.
fraunhofer.de

†ORCID:

Charles Tapley Hoyt
orcid.org/0000-0003-4423-4370

Specialty section:

This article was submitted to
Bioinformatics and
Computational Biology,
a section of the journal
Frontiers in Genetics

Received: 23 August 2019

Accepted: 30 October 2019

Published: 22 November 2019

Citation:

Mubeen S, Hoyt CT, Gemünd A,
Hofmann-Apitius M, Fröhlich H and
Domingo-Fernández D (2019) The
Impact of Pathway Database Choice
on Statistical Enrichment Analysis
and Predictive Modeling.
Front. Genet. 10:1203.
doi: 10.3389/fgene.2019.01203

Pathway-centric approaches are widely used to interpret and contextualize -omics data. However, databases contain different representations of the same biological pathway, which may lead to different results of statistical enrichment analysis and predictive models in the context of precision medicine. We have performed an in-depth benchmarking of the impact of pathway database choice on statistical enrichment analysis and predictive modeling. We analyzed five cancer datasets using three major pathway databases and developed an approach to merge several databases into a single integrative one: MPath. Our results show that equivalent pathways from different databases yield disparate results in statistical enrichment analysis. Moreover, we observed a significant dataset-dependent impact on the performance of machine learning models on different prediction tasks. In some cases, MPath significantly improved prediction performance and also reduced the variance of prediction performances. Furthermore, MPath yielded more consistent and biologically plausible results in statistical enrichment analyses. In summary, this benchmarking study demonstrates that pathway database choice can influence the results of statistical enrichment analysis and predictive modeling. Therefore, we recommend the use of multiple pathway databases or integrative ones.

Keywords: pathway enrichment, benchmarking, databases, machine learning, statistical hypothesis testing

INTRODUCTION

As fundamental interactions within complex biological systems have been discovered in experimental biology labs, they have often been assembled into computable pathway representations. Because they have proven immensely useful in the analysis and interpretation of -omics data when coupled with algorithmic approaches (e.g., gene set enrichment analysis, GSEA), academic and commercial groups have generated and maintained a comprehensive set of databases during the last 15 years (Bader et al., 2006). Examples include KEGG, Reactome, WikiPathways, NCIPathways, and Pathway Commons (Schaefer et al., 2008; Cerami et al., 2011; Kanehisa et al., 2016; Slenter et al., 2017; Fabregat et al., 2018).

However, these databases tend to differ in the average number of pathways they contain, the average number of proteins per pathway, the types of biochemical interactions they incorporate, and the subcategories of pathways that they provide (e.g., signal transduction, genetic interaction, and metabolic) (Kirouac et al., 2012; Türei et al., 2016). Pathways are often also described at varying levels of detail, with diverse data types and with loosely defined boundaries (Domingo-Fernández et al.,

2018). Nonetheless, most pathway analyses are still conducted exclusively by employing a single database, often chosen in part by researchers' preferences or previous experiences (e.g., bias towards a database previously yielding good results and ease of use of a particular database) (Table 1). Notably, the selection of a suitable pathway database depends on the actual biological context that is investigated, yet KEGG remains severely overrepresented in published -omics studies. This raises concerns and motivates the consideration of multiple pathway databases or, preferably, an integration over several pathways resources.

Several integrative resources have been developed, including meta-databases [e.g., Pathway Commons (Cerami et al., 2011), MSigDB (Liberzon et al., 2015), and ConsensusPathDB (Kamburov et al., 2008)] that enable pathway exploration in their corresponding web applications and integrative software tools [e.g., graphite (Sales et al., 2018), PathMe (Domingo-Fernandez et al., 2019), and OmniPath (Türei et al., 2016)] designed to enable bioinformatics analyses. By consolidating pathway databases, these resources have attempted to summarize major reference points in the existing knowledge and demonstrate how data contained in one resource can be complemented by data contained in others. Thus, through their usage, the biomedical community has benefitted from comprehensive overviews of pathway landscapes which can then make for more robust resources highly suited for analytic usage.

The typical approach to combine pathway information with -omics data is *via* statistical enrichment analysis, also known as pathway enrichment. The task of navigating through the continuously developing variants of enrichment methods has been undertaken by several recent studies which benchmarked the performance of these techniques (Bayerlová et al., 2015; Ilnatova et al., 2018; Lim et al., 2018) and guide users on the choice for their analyses (Fabris et al., 2019; Reimand et al., 2019). While Bateman et al. (2014) examined the impact of choice of different subsets of MSigDB on GSEA, it remains unclear what broader impact an integrative pathway meta-database would have for statistical enrichment analysis. Additionally, the overlap of pathways within the same integrative database can induce biases (Liberzon et al., 2015), specifically when conducting multiple testing correction *via* the popular Benjamini–Hochberg method (Benjamini and Hochberg, 1995) that supposes independence of statistical tests. This issue is of particular concern for large-scale meta-databases such as MSigDB.

The aim of this work is to systematically investigate the influence of alternative representations of the same biological pathway (e.g., in KEGG, Reactome, and WikiPathways) on the results of statistical enrichment analysis *via* three common methods: the hypergeometric test, GSEA, and signaling pathway impact analysis (SPIA) (Fisher, 1992; Subramanian et al., 2005; Tarca et al., 2008) using five The Cancer Genome Atlas (TCGA) datasets (Weinstein et al., 2013). In addition, we also show that pathway activity-based patient classification and survival analysis *via* single-sample GSEA (ssGSEA; Barbie et al., 2009) can be impacted by the choice of pathway resource in some cases. As a solution, we propose to integrate different pathway resources *via* a method where semantically analogous pathways across databases (e.g., "Notch signaling pathway" in KEGG and "Signaling by NOTCH" pathway in Reactome) are combined. This approach exploits the pathway mappings and harmonized pathway representations described in our previous work (Domingo-Fernández et al., 2018; Domingo-Fernandez et al., 2019). We demonstrate that when aided by our integrative pathway database, it is possible to better capture expected disease biology than with individual resources, and to sometimes obtain better predictions of clinical endpoints. Our entire analytic pipeline is implemented in a reusable Python package (`pathway_forte`; see *Materials and Methods*) to facilitate reproducing the results with other databases or datasets in the future.

MATERIALS AND METHODS

In the first two subsections, we describe the pathway resources and the clinical and genomic datasets we used in benchmarking. The following sections then outline the statistical enrichment analysis and predictive modeling conducted in this study. Finally, in the last two subsections, we describe the statistical methods and the software implemented to conduct the benchmarking.

Pathway Databases

Selection Criteria

Numerous viable pathway databases have been made available to infer biologically relevant pathway activity (Bader et al., 2006). In this work, we systematically compared three major ones (i.e., KEGG, Reactome, and WikiPathways) as the subset of databases to benchmark. The rationale for the inclusion of these databases was twofold: firstly, these databases are open-sourced, well-established, and highly cited in studies investigating pathways associated with variable gene expression patterns in different sets of conditions (Table 1). Secondly, we expected distinctions between these databases to be strong enough to observe variable results of enrichment analysis and patient classification, yet these databases also contain a reasonable number of equivalent pathways such that objective comparisons could be made, as outlined in our previous work (Domingo-Fernández et al., 2018).

Data Retrieval and Processing

In order to systematically compare results yielded by different databases, we retrieved the contents of KEGG, Reactome, and WikiPathways using ComPath (Domingo-Fernández et al., 2018)

TABLE 1 | Number of publications citing major pathway resources for pathway enrichment in PubMed Central (PMC), 2019. To develop an estimate on the number of publications using several pathway databases for pathway enrichment, SCAIView (<http://academia.scaiview.com/academia>; indexed on 01/03/2019) was used to conduct the following query using the PMC corpus: "<pathway resource>" AND "pathway enrichment".

Type	Pathway resource	Publications
Primary	KEGG	27,713
	Reactome	3,765
	WikiPathways	651
Integrative	MSigDB	2,892
	ConsensusPathDB	339
	Pathway Commons	1,640

and converted it into the Gene Matrix Transposed (GMT) file format. Generated networks encoded in Biological Expression Language (BEL; Slater, 2014) were retrieved using PathMe (Domingo-Fernández et al., 2019).

To test the potential utility of an integrative pathway resource, we used equivalent pathways across the three databases that were manually curated in our previous work (Domingo-Fernández et al., 2018; see our earlier publication for further details). In the following, we call these “pathways analogs” or “equivalent pathways” (Figure 1A), while we call a pathway found as analogous across all KEGG, Reactome, as well as WikiPathways a “super pathway”.

In a second step, we merged equivalent pathways by taking the graph union with respect to contained genes and interactions (Figures 1B, C). We have also described this step in more detail in our earlier work (Domingo-Fernández et al., 2019).

The set union of KEGG, Reactome, and WikiPathways, while taking into account pathway equivalence, gave rise to an integrative resource to which we refer as *MPath* (Figure 1D). By merging equivalent pathways, *MPath* contains a fewer number of pathways than the sum of all pathways from all primary resources. In total, *MPath* contains 2,896 pathways, of which 238 are derived from KEGG, 2,119 from Reactome, and 409 from

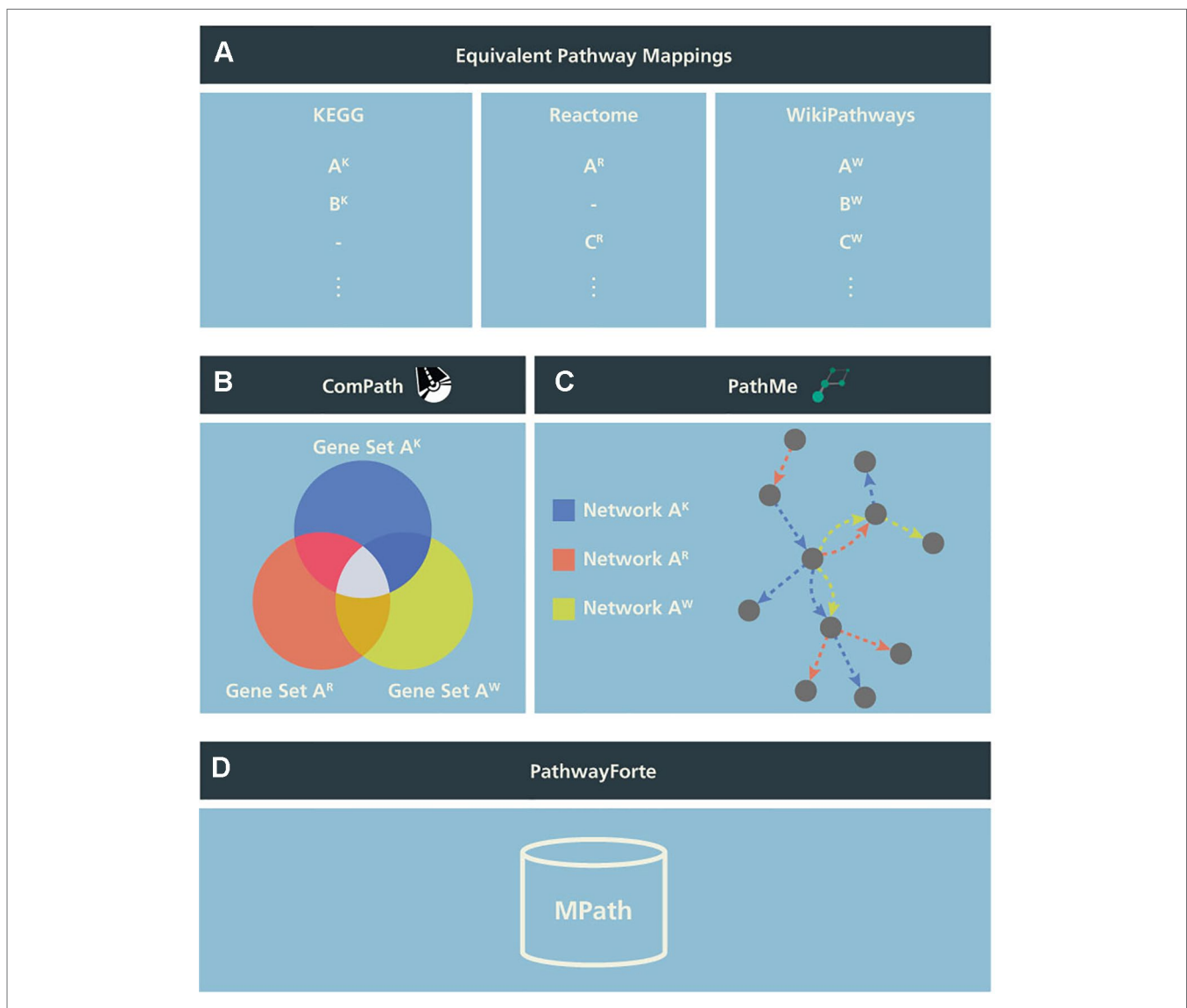


FIGURE 1 | Schema illustrating the generation of *MPath*. The curated pathway mapping catalog is depicted in (A), which links equivalent pathways from different resources. Pathways that are shared across two resources are referred to as pathway analogs (i.e., Pathway A in Reactome and Pathway A' in KEGG) and pathways that are shared across all three resources are referred to as “super pathways” (i.e., Pathway A in KEGG, Pathway A' in Reactome, and Pathway A'' in WikiPathways). (B) Using these mappings, gene sets of equivalent pathways from different resources can be combined, ensuring key molecular players from the different resources are included. (C) Similarly, network representations of the pathways can be overlaid to generate more comprehensive pathways. (D) Finally, both the combined gene sets and networks representations are included in *MPath*. Note that pathways that are exclusive to a single database are included in *MPath* unchanged.

WikiPathways, while another 129 pathways are pathway analogs and 26 are super pathways.

We next compared the latest versions of pathway gene sets from KEGG, Reactome, WikiPathways, and MPath with pathway gene sets from MSigDB, a highly cited integrative pathway database containing older versions of the KEGG and Reactome gene sets (Liberzon et al., 2015). We downloaded KEGG and Reactome gene sets from the curated gene set (C2) collection of MSigDB (<http://software.broadinstitute.org/gsea/msigdb/collections.jsp#C2>; version 6.2; July 2018). Detailed statistics on the number of pathways from each resource are presented in **Table S1**.

Clinical and Genomic Data

We used five widely used datasets acquired from TCGA (Weinstein et al., 2013), a cancer genomics project that has catalogued molecular and clinical information for normal and tumor samples (**Table 2**). TCGA data were retrieved through the Genomic Data Commons (GDC; <https://gdc.cancer.gov>) portal and cBioportal (<https://www.cbioportal.org>) on 14-03-2019. RNA-seq gene expression data subjected to an mRNA quantification analysis pipeline for BRCA, KIRC, LIHC, OV, and PRAD TCGA datasets were queried, downloaded, and prepared from the GDC through the R/Bioconductor package, TCGAbiolinks (R version: 3.5.2; TCGAbiolinks version: 2.10.3) (Colaprico et al., 2015). The data were preprocessed as follows: gene expression was quantified by the number of reads aligned to each gene and read counts were measured using HTSeq and normalized using fragments per kilobase of transcript per million mapped reads upper quartile (FPKM-UQ). HTSeq raw read counts also subject to the GDC pipeline were similarly queried, downloaded, and prepared with TCGAbiolinks. Read count data downloaded for the BRCA, KIRC, LIHC, and PRAD datasets were processed to remove identical entries, while unique measurements of identical genes were averaged. The differential gene expression analysis of cancer versus normal samples was performed using the R/Bioconductor package, DESeq2 (version 1.22.2). Genes with adjusted p value < 5% were considered significantly dysregulated. For all downloaded data, gene identifiers were mapped to HGNC gene symbols (Povey et al., 2001), where possible. To obtain additional information on the survival status and time to death, or censored survival times of patients, patient identifiers in the TCGA datasets were mapped to their equivalent identifiers in cBioPortal. Additionally, cancer subtype classifications or the PRAD and

BRCA datasets were retrieved from the GDC. We would like to note that although there are other cohorts available (e.g., COAD and STAD) containing all of these modalities, we did not include them in this analysis because of the limited number of samples they contain (i.e., less than 300 patients). Detailed statistics of all five datasets are presented in **Table 2**.

Pathway Enrichment Methods

In this subsection, we describe three different classes of pathway enrichment methods that we tested: 1) statistical overrepresentation analysis (ORA); 2) functional class scoring (FCS); and 3) pathway topology (PT)-based enrichment (**Figure 2**) (Khatri et al., 2012; García-Campos et al., 2015; Fabris et al., 2019).

Overrepresentation Analysis

We conducted pathway enrichment using genes that exhibited a q value < 0.05 using a one-sided Fisher's exact test (Fisher, 1992) for each of the pathways in all pathway databases. We consider a pathway to be significantly enriched if its q value is smaller than 0.05 after applying multiple hypothesis testing correction with the Benjamini–Yekutieli method under dependency (Benjamini and Yekutieli, 2001).

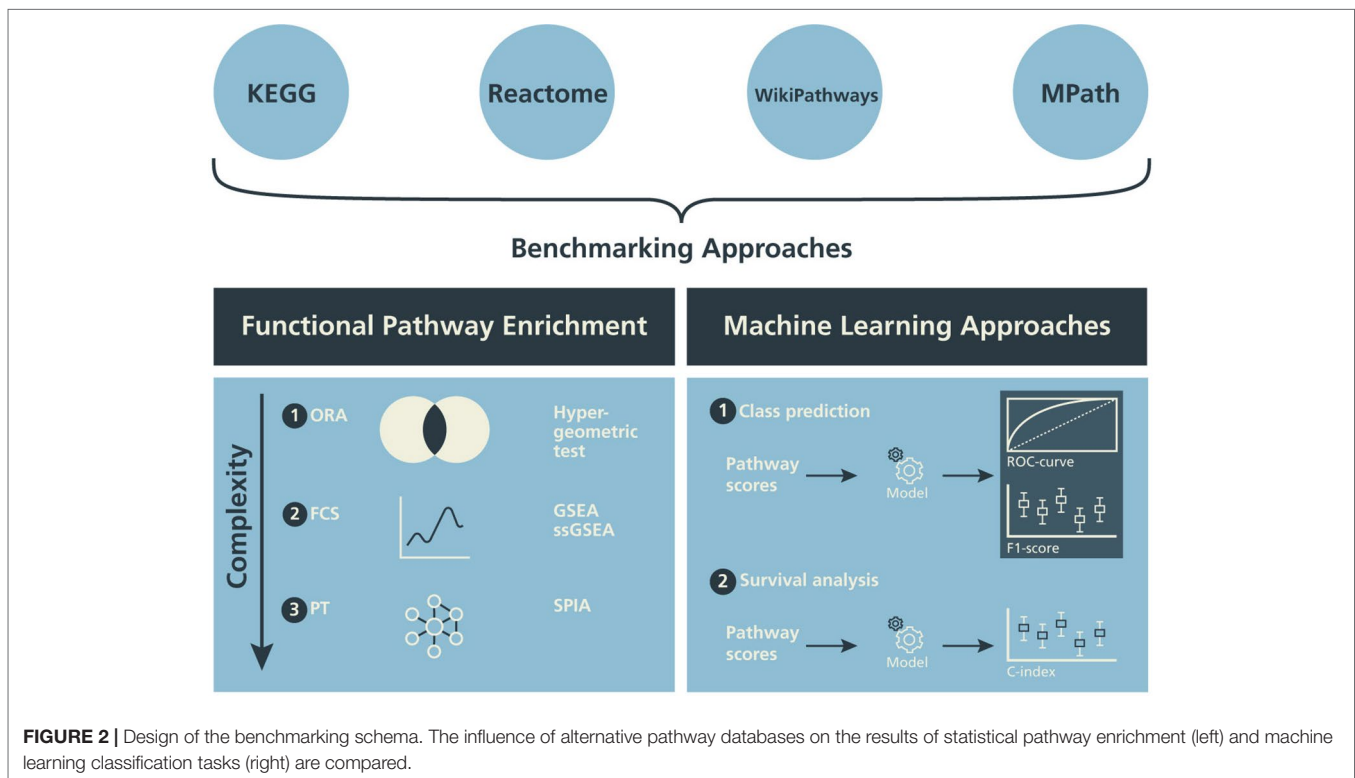
Functional Class Scoring Methods

We selected GSEA, one of the most commonly used FCS methods (Subramanian et al., 2005). We performed GSEA with the Python package, GSEAPy (version 0.9.12; <https://github.com/zqfang/gseapy>), using normalized RNA-seq expression quantifications (FPKM-UQ) obtained for the BRCA, KIRC, LIHC, and PRAD datasets containing both normal and tumor samples (**Table 2**). All genes were ranked by their differential expression based on their \log_2 fold changes. Query gene sets for GSEA included pathways from KEGG, Reactome, WikiPathways, and MPath. GSEA results were filtered to include pathway gene sets with p values below 0.05 and a minimum gene set size of 10 or a maximum gene size of 3,000. Similarly, GSEAPy was used to perform ssGSEA (Barbie et al., 2009) (**Table S2**) to acquire sample-wise pathway scores using FPKM-UQ for BRCA, KIRC, LIHC, OV, and PRAD datasets, irrespective of phenotype labels (Barbie et al., 2009). Datasets were filtered to only include normalized expression data for genes found in the pathway gene sets of KEGG, Reactome, WikiPathways, and MPath and then used for ssGSEA. Expression data were ranked and sample-wise normalized enrichment scores were obtained.

TABLE 2 | Statistics of the five TCGA cancer datasets used in this work.

Cancer type	TCGA abbreviation	Tumor samples	Normal samples	Surviving patients	Deceased patients
Breast invasive carcinoma	BRCA	1,102	113	946	153
Kidney renal clear cell carcinoma	KIRC	538	72	365	173
Liver hepatocellular carcinoma	LIHC	371	50	240	130
Prostate adenocarcinoma	PRAD	498	52	498	10
Ovarian cancer	OV	374	0	143	229

The statistics correspond to those retrieved from the GDC portal and cBioportal on 14-03-2019. Longitudinal statistics of survival data are presented in *Figure S1*.



Pathway Topology-Based Enrichment

To evaluate PT-based methods, we selected the well-known and highly cited SPIA method (Tarca et al., 2008) for two main reasons: firstly, the guidelines outlined by a comparative study on topology-based methods (Ihnatova et al., 2018) recommend the use of SPIA for datasets with properties similar to TCGA (i.e., possessing two well-defined classes, full expression profiles, many samples, and numerous differentially expressed genes). Secondly, SPIA has been reported to have a high specificity while preserving dependency on topological information (Ihnatova et al., 2018). Because the R/Bioconductor's SPIA package only contains KEGG pathways, we converted the pathway topologies from the three databases used in this work to a custom format in a similar fashion as graphite (Sales et al., 2018) (**Supplementary Text**). We declared significance for SPIA-based pathway enrichment, if the Bonferroni corrected p value was $<5\%$.

Evaluation Based on Enrichment of Pathway Analogs

To better understand the impact of database choice, we compared the raw p value rankings (i.e., before multiple testing correction) of pathway analogs across each possible pair of databases (i.e., in KEGG and Reactome, Reactome and WikiPathways, and WikiPathways and KEGG) and in each statistical enrichment analysis (i.e., hypergeometric test, GSEA, and SPIA) with the Wilcoxon signed-rank test. It assessed the average rank difference of the pathway analogs and reported how significantly different the results were for each database pair. Importantly, we only tested statistical enrichment of the analogous pathways in order to avoid statistical biases due to differences in the size of pathway databases.

Machine Learning

ssGSEA was conducted to summarize the gene expression profile mapping to a particular pathway of interest within a given patient sample, hence resulting in a pathway activity profile for each patient. We then evaluated the different pathway resources with respect to three machine learning tasks:

1. Prediction of tumor vs. normal
2. Prediction of known tumor subtype
3. Prediction of overall survival

Prediction of Tumor vs. Normal

The first task was to train and evaluate binary classifiers to predict normal versus tumor sample labels. This task was conducted for four of the five TCGA datasets (i.e., BRCA, KIRC, LIHC, and PRAD), while OV, which only contains tumor samples, was omitted. We performed this classification using a commonly used elastic net penalized logistic regression model (Zou and Trevor, 2005). Prediction performance was evaluated *via* a 10 times repeated 10-fold stratified cross-validation. Importantly, tuning of elastic net hyper-parameters (l_1 , l_2 regularization parameters) was conducted within the cross-validation loop to avoid over-optimism (Molinari et al., 2005).

Prediction of Tumor Subtype

The second task was to train and evaluate multi-label classifiers to predict tumor subtypes using sample-wise pathway activity scores generated from ssGSEA. This task was only conducted for the BRCA and PRAD datasets, similar to the work done by Lim et al. (2018), because the remaining three datasets included

in this work lacked subtype information. From the five breast cancer subtypes present in the BRCA dataset by the PAM50 classification method (Sorlie et al., 2001), we included four subtypes (i.e., 194 Basal samples, 82 Her2 samples, 567 LumA samples, and 207 LumB samples). These four were selected as they constitute the agreed-upon intrinsic breast cancer subtypes according to the 2015 St. Gallen Consensus Conference (Coates et al., 2015) and are also recommended by the ESMO Clinical Practice Guidelines (Senkus et al., 2015). For the PRAD dataset, evaluated subtypes included 151 ERG samples, 27 ETV1 samples, 14 ETV4 samples, 38 SPOP samples, and 87 samples classified as other (Cancer Genome Atlas Research Network, 2014). Similar to the approach by Graudenzi et al. (2017), support vector machines (SVMs) (Cortes and Vapnik, 1995) were used for subtype classification by implementing a one-versus-one strategy in which a single classifier is fit for each pair of class labels. This strategy transforms a multi-class classification problem into a set of binary classification problems. We again used a 10 times repeated 10-fold cross-validation scheme, and the soft margin parameter of the linear SVM was tuned within the cross-validation loop *via* a grid search. We assessed the multi-class classifier performance in terms of accuracy, precision, and recall.

Prediction of Overall Survival

The third task was to train and evaluate machine learning models to predict overall survival of cancer patients. For this purpose, a Cox proportional hazards model with elastic net penalty was used (Tibshirani, 1997; Friedman et al., 2010). Prediction performance was evaluated on the basis of five TCGA datasets (i.e., BRCA, LIHC, KIRC, OV, and PRAD) (Table 2) using the same 10 times repeated 10-fold nested cross-validation procedure as described before. The performance of the model was assessed by Harrell's concordance index (c-index; Harrell et al., 1982), which is an extension of the well-known area under receiver operating characteristic (ROC) curve for right censored time-to-event (here: death) data.

Statistical Assessment of Database Impact on Prediction Performance

To understand the degree to which the observed variability of area under the ROC curve (AUC) values, accuracies, and c-indices could be explained by the actually used pathway resource, we conducted a two-way analysis of variance (ANOVA). The ANOVA model had the following form:

$$\text{performance} \sim \text{database} + \text{dataset} + \text{database} \times \text{dataset}$$

We then tested the significance of the database factor *via* an *F* test. In addition, we performed Wilcoxon tests analysis to understand specific differences between databases in a dataset-dependent manner.

Software Implementation

The workflow presented in this article consists of three major components: 1) the acquisition and preprocessing of gene set

and pathway databases; 2) the acquisition and preprocessing of experimental datasets; and 3) the re-implementation or adaptation of existing analytical pipelines for benchmarking. We implemented these components in the `pathway_forte` Python package to facilitate the reproducibility of this work, the inclusion of additional gene set and pathway databases, and to include additional experimental datasets.

The acquisition of KEGG, MSigDB, Reactome, and WikiPathways was mediated by their corresponding Bio2BEL Python packages (Hoyt et al., 2019; <https://github.com/bio2bel>) in order to provide uniform access to the underlying databases and to enable the reproduction of this work as they are updated. Each Bio2BEL package uses Python's *entry points* to integrate in the previously mentioned ComPath framework in order to support uniform preprocessing and enable the integration of further pathway databases in the future, without changing any underlying code in the `pathway_forte` package. The network preprocessing defers to PathMe (Domingo-Fernandez et al., 2019; <https://github.com/pathwaymerger>). Because it is based on PyBEL (Hoyt et al., 2018; <https://github.com/pybel>), it is extensible to the growing ecosystem of BEL-aware software.

While the acquisition and preprocessing of experimental datasets is currently limited to a subset of TCGA, it is extensible to further cancer-specific and other condition-specific datasets. We implemented independent preprocessing pipelines for several previously mentioned datasets using extensive manual curation, preparation, and processing with the `pandas` Python package (McKinney, 2010; <https://github.com/pandas-dev/pandas>). Unlike the pathway databases, which were amenable to standardization, the preprocessing of each new dataset must be bespoke.

The re-implementation and adaptation of existing analytical methods for functional enrichment and prediction involved wrapping several existing analytical packages (Table S3) in order to make their application programming interfaces more user-friendly and to make the business logic of the benchmarking more elegantly reflected in the source code of `pathway_forte`. Each is independent and can be used with any combination of pathway database and dataset. Finally, all figures presented in this paper and complementary analyses can be generated and reproduced with the Jupyter notebooks located at <https://github.com/pathwayforte/results/>.

Ultimately, we wrapped each of these components in a command line interface (CLI) such that the results presented in each section of this work can be generated with a corresponding command following the guidelines described by Grüning et al. (2019). The scripts for generating the figures in this manuscript are not included in the main `pathway_forte`, but rather in their own repository within Jupyter notebooks at <https://github.com/PathwayForte/results>.

The source code of the `pathway_forte` Python package is available at <https://github.com/PathwayForte/pathway-forte>, its latest documentation can be found at <https://pathwayforte.readthedocs.io>, and its distributions can be found on PyPI at <https://pypi.org/project/pathway-forte>.

The `pathway_forte` Python package has a tool chain consisting of `pytest` (<https://github.com/pytest-dev/pytest>) as a testing

framework, coverage (<https://github.com/nedbat/coveragepy>) to assess testing coverage, sphinx (<https://github.com/sphinx-doc/sphinx>) to build documentation, flake8 (<https://github.com/PyCQA/flake8>) to enforce code and documentation quality, setuptools (<https://github.com/pypa/setuptools>) to build distributions, pyroma (<https://github.com/regebro/pyroma>) to enforce package metadata standards, and tox (<https://github.com/tox-dev/tox>) as a build tool to facilitate the usage of each of these tools in a reproducible way. It leverages community and open-source resources to improve its usability by using Travis-CI (<https://travis-ci.com>) as a continuous integration service, monitoring testing coverage with Codecov (<https://codecov.io>), and hosting its documentation on Read the Docs (<https://readthedocs.org>).

Hardware

Computations for each of the tasks were performed on a symmetric multiprocessing (SMP) node with four Intel Xeon Platinum 8160 processors per node with 24 cores/48 threads each (96 cores/192 threads per node in total) and 2.1-GHz base/3.7-GHz Turbo Frequency with 1,536-GB/1.5-TB RAM (DDR4 ECC Reg). The network was 100 GBit/s Intel OmniPath, storage was 2× Intel P4600 1.6-TB U.2 PCIe NVMe for local intermediate data and BeeGFS parallel file system for Home directories. **Table 3** provides a qualitative description of the memory and time requirements for each task.

RESULTS

The results of the benchmarking study have been divided into two subsections for each of the pathway methods described above. We first compared the effects of database selection on the results of functional pathway enrichment methods. In the following subsection, we benchmarked the performance of the pathway resources on the various machine learning classification tasks conducted.

Benchmarking the Impact on Enrichment Methods

Overrepresentation Analysis

As illustrated by our results, pathway analogs from different pathway databases in several cases showed clearly significant

rank differences (**Figure 3**). These differences were most pronounced between Reactome and WikiPathways. For example, while the "Thyroxine Biosynthesis" pathway was highly statistically significant (q value <0.01) in the LIHC dataset for Reactome, its analogs in WikiPathways (i.e., "Thyroxine (Thyroid Hormone) Production") and KEGG (i.e., "Thyroid Hormone Synthesis") were not. However, the pathway was found to be significantly enriched in MPath. Such differences were similarly observed for the "Notch signaling" pathway in the PRAD dataset, in which the pathway was highly statistically significant (q value <0.01) for Reactome and MPath, but showed no statistical significance for KEGG and WikiPathways. Similar cases were systematically observed for additional pathway analogs and super pathways, demonstrating that marked differences in rankings can arise depending on the database used.

Gene Set Enrichment Analysis

Similar to ORA, GSEA showed significant differences between pathway analogs across databases in several cases (**Figure 3**). These differences were most pronounced between KEGG and WikiPathways in the KIRC and LIHC datasets and between KEGG and Reactome in the BRCA and PRAD datasets. Since GSEA calculates the observed direction of regulation (e.g., over/underexpressed) of each pathway, we also examined whether super pathways or pathway analogs exhibited opposite signs in their normalized enrichment scores (NES) (e.g., one pathway is overexpressed while its equivalent pair is underexpressed). As an illustration, GSEA results of the LIHC dataset revealed the contradiction that the "DNA replication" pathway, one of 26 super pathways, was overexpressed according to Reactome and underexpressed according to KEGG and WikiPathways, though the pathway was not statistically significant for any of these databases. However, the merged "DNA replication" pathway in MPath appeared as significantly underexpressed. Similarly, in the BRCA dataset, the WikiPathways definition of the "Notch signaling" and "Hedgehog signaling" pathways were significantly overexpressed, while the KEGG and Reactome definitions were insignificantly overexpressed. Interestingly, both the merged "Notch signaling" and merged "Hedgehog signaling" pathways appeared as significantly underexpressed ($q < 0.05$) in MPath.

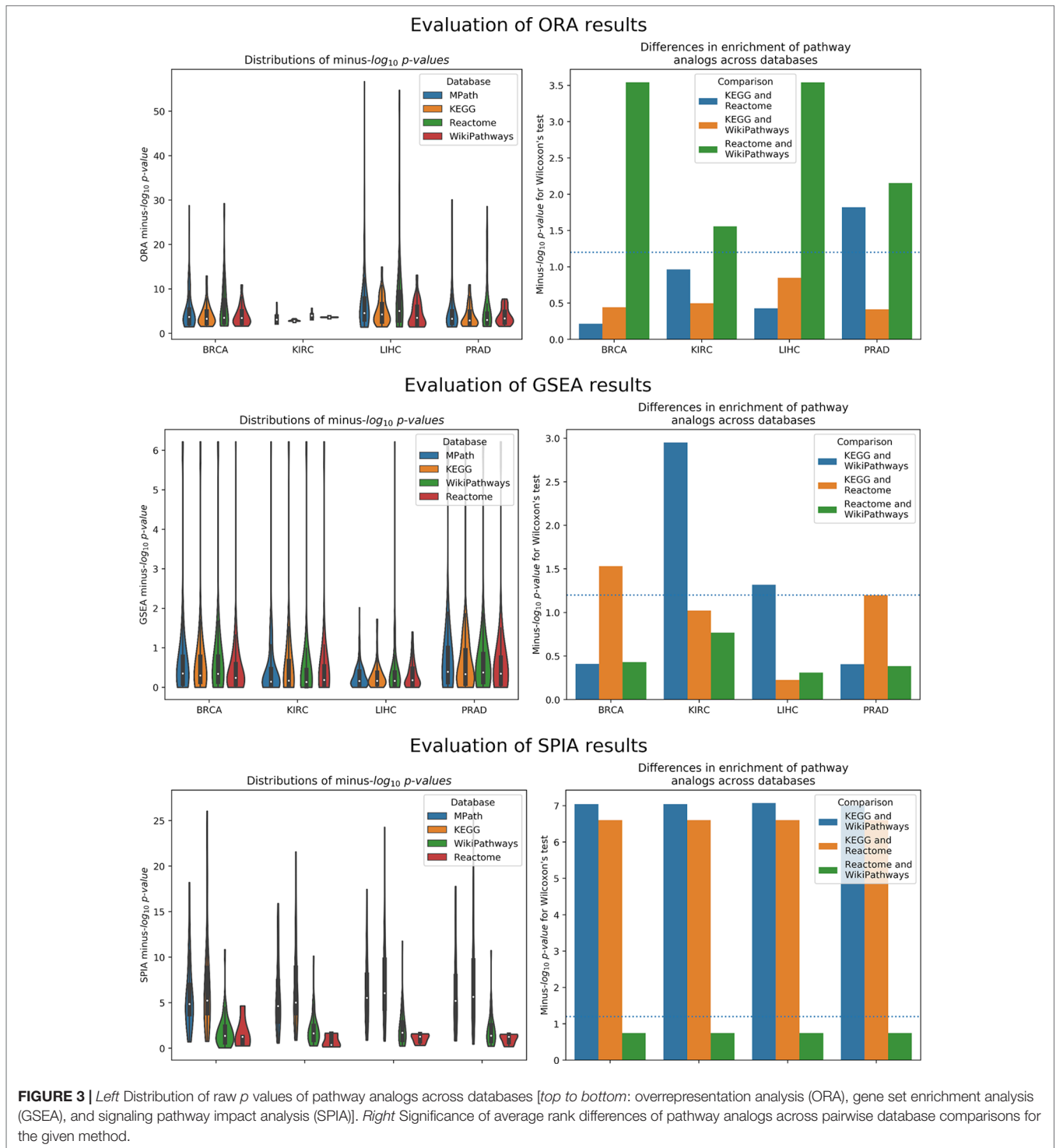
Signaling Pathway Impact Analysis

The final of the three statistical enrichment analyses conducted revealed further differences between pathway analogs across databases. As expected, differences in the results of analogous pathways were exacerbated on topology-based methods compared with ORA and GSEA, as these latter methods do not consider pathway topology (i.e., incorporation of pathway topology introduces one extra level of complexity, leading to higher variability) (**Figure 3**). Beyond a cursory inspection of the statistical results, we also investigated the concordance of the direction of change of pathway activity (i.e., activation or inhibition) for equivalent pathways. We found that for two database (i.e., LIHC and KIRC), the direction of change was inconsistently reported for the "TGF beta signaling" pathway, depending on the database used (i.e., the KEGG representation

TABLE 3 | A qualitative description of the computational costs of the analyses performed.

Task	Relative memory usage	Timescale
ORA	Low	Seconds
GSEA	Medium	Minutes
ssGSEA	Very high	Hours
Prediction of tumor vs. normal	Medium	Minutes
Prediction of known tumor subtype	Medium	Minutes
Prediction of overall survival	Medium	Hours

Performing ssGSEA required on the scale of 100 GB of RAM for some dataset/database combinations, while the other tasks could be run on a modern laptop with no issues.

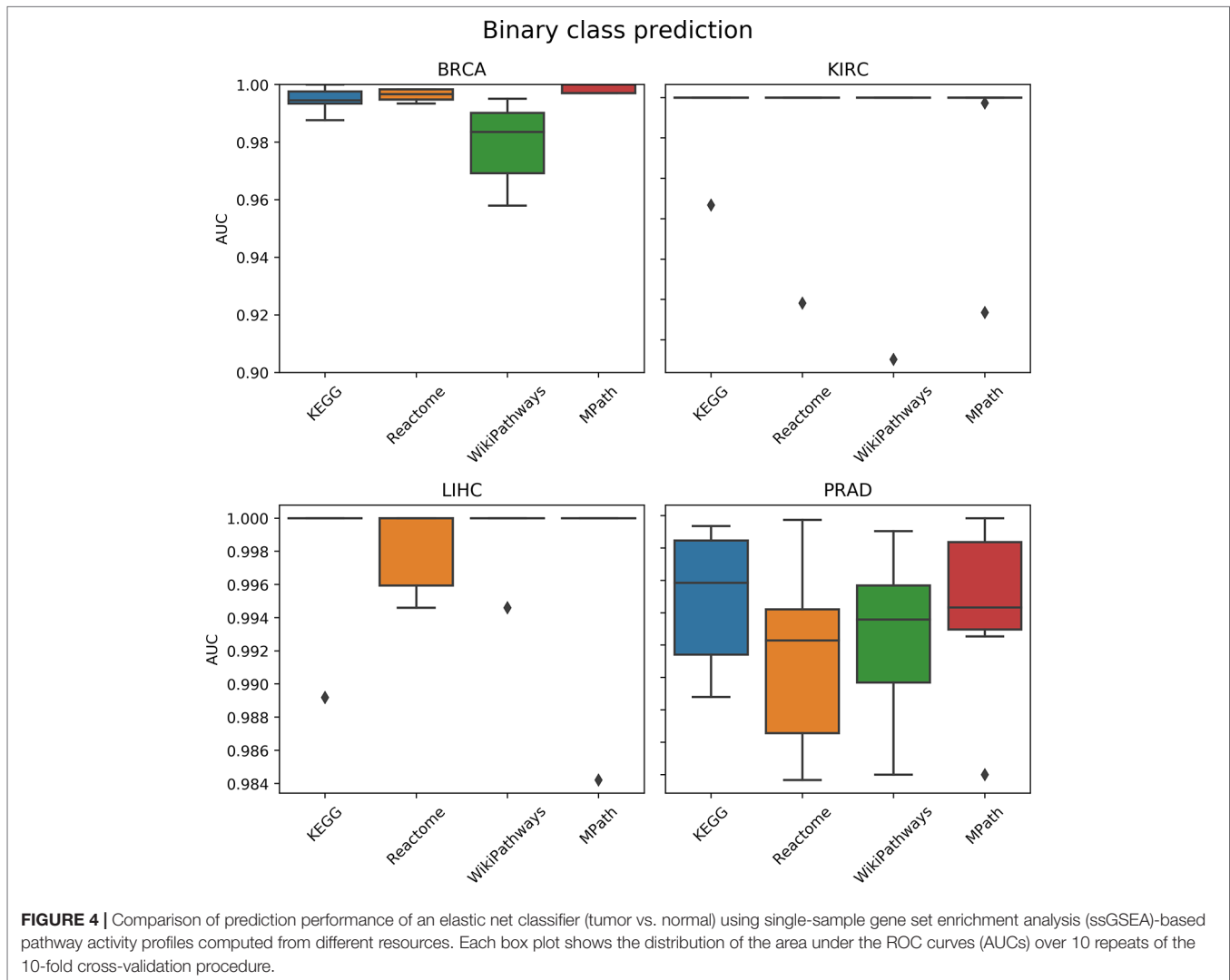


was activated and the WikiPathways one inhibited). A similar effect was observed in the "Estrogen signaling pathway," found to be inhibited in KEGG and activated in WikiPathways in the LIHC dataset. The merging of equivalent pathway networks resulted in the observation of inhibition for both the "TGF beta signaling" and "Estrogen signaling" pathways in MPath results.

Benchmarking the Impact on Predictive Modeling

Prediction of Tumor vs. Normal

We compared the prediction performance of an elastic net penalized logistic regression classifier to discriminate normal from cancer samples based on their pathway activity profiles. The cross-validated prediction performance was measured



via the AUC and precision-recall curve (see the corresponding *Materials and Methods* section). The AUC indicated no overall significant effect of the choice of pathway database on model prediction performance ($p = 0.5$, ANOVA F test; **Figure 4**). Similarly, the results of the precision-recall curve did not show a significant effect of the database selected on the model's predictive performance. Finally, these results were not surprising due to the relative ease of the classification task (i.e., all AUC values were close to 1).

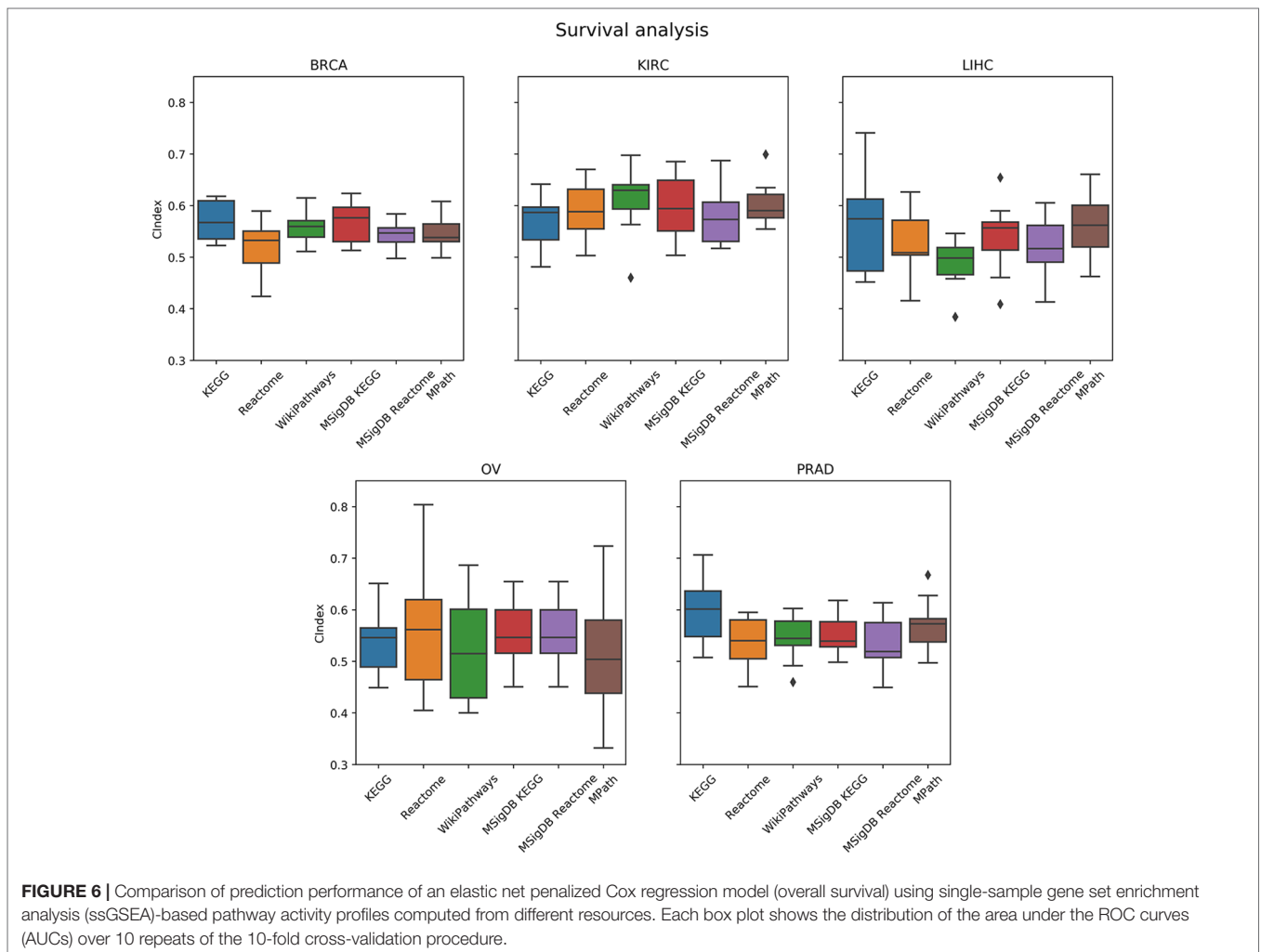
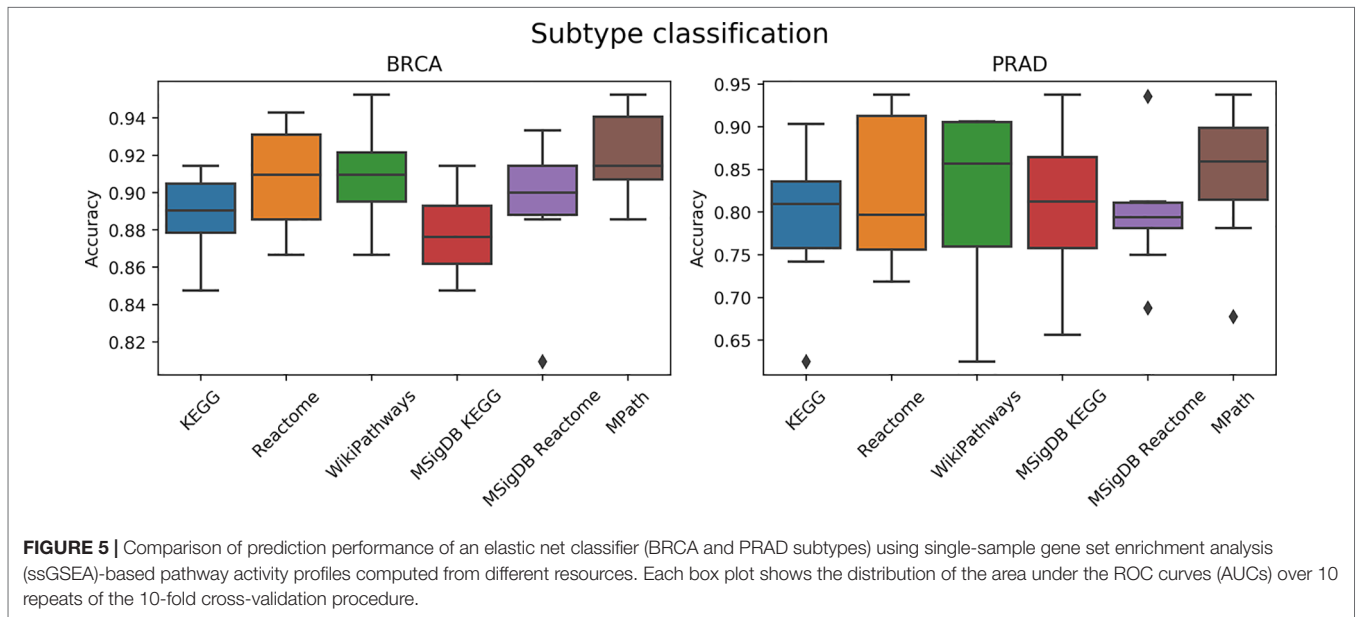
Prediction of Tumor Subtype

We next compared the prediction performances of a multi-class classifier predicting known tumor subtypes of BRCA and PRAD using ssGSEA-based pathway activity profiles. **Figure 5** demonstrated no overall significant effect of the choice of pathway database ($p = 0.16$, ANOVA F test). We used Wilcoxon tests to investigate if each pair of distributions of the accuracies based on each database were different, but did

not achieve statistical significance ($q < 0.01$) after Benjamini-Hochberg correction for multiple hypothesis testing. While the lack of significance is probably due to the limited amount of datasets (only two contained subtype information) and measurements, we would like to note that MPath showed the best classification metrics (similar to the previous classification task).

Prediction of Overall Survival

As a next step, we compared the prediction performance of an elastic net penalized Cox regression model for overall survival using ssGSEA-based pathway activity profiles derived from different resources. As indicated in **Figure 6**, no overall significant effect of the actually used pathway database could be observed ($p = 0.28$, ANOVA F test). A limiting factor of this analysis is the fact that overall survival can generally only be predicted slightly above chance level (c-indices range between 55% and 60%) based on gene expression alone, which is in agreement with the



literature (Van Wieringen et al., 2009; Fröhlich, 2014; Mayr and Schmid, 2014; Zhang et al., 2018).

DISCUSSION

In this work, we presented a comprehensive comparative study of pathway databases based on functional enrichment and predictive modeling. We have shown that the choice of pathway database can significantly influence the results of statistical enrichment, which raises concerns about the typical lack of consideration that is given to the choice of pathway resource in many gene expression studies. This finding was specifically pronounced for SPIA because this method is a topology-based enrichment approach and therefore expected to be most sensitive to the actual definition of a pathway. At the same time, we observed that an integrative pathway resource (MPath) led to more biologically consistent results and, in some cases, improved prediction performance.

Generating a merged dataset such as MPath is non-trivial. We purposely restricted this study to three major pathway databases because of the availability of inter-database pathway mappings and pathway networks from our previous work which enabled conducting objective database comparisons. The incorporation of additional pathway databases into MPath would first require the curation of pathway mappings prior to conducting the benchmarking study, which can be labor-intensive. Furthermore, performing the tasks described in this work comes with a high computational cost (Table 1).

Our strategy to build MPath is one of many possible approaches to integrate pathway knowledge from multiple databases. Although alternative meta-databases such as Pathway Commons and MSigDB do exist, the novelty of this work lies in the usage of mappings and harmonized pathway representations for generating a merged dataset. While we have presented MPath as one possible integrative approach, alternative meta-databases may be used, but would require that researchers ensure that the meta-databases' contents are continuously updated (Wadi et al., 2016).

Our developed mapping strategy between different graph representations of analogous pathways enabled us to objectively compare pathway enrichment results that otherwise would have been conducted manually and subjectively. Furthermore, they allowed us to generate super pathways inspired by previous approaches that have shown the benefit of merging similar pathway representations (Doderer et al., 2012; Vivar et al., 2013; Belinky et al., 2015; Stoney et al., 2018; Miller et al., 2019). In this case, this was made possible by the fully harmonized gene sets and networks generated by our previous work, ComPath and PathMe. A detailed description of the ComPath and PathMe publications, source code, and extensions to existing analyses (i.e., SPIA) to better suit the methods used in this work can be found in the **Supplementary Text**.

One of the limitations of this work is that we restricted the analysis to five cancer datasets from TCGA and we did

not expand it to other conditions besides cancer. The use of this disease area was mainly driven by the availability of data and the corresponding possibilities to draw statistically valid conclusions. However, we acknowledge the fact that data from other disease areas may result in different findings. More specifically, we believe that a similar benchmarking study based on data from disease conditions with an unknown pathophysiology (e.g., neurological disorders) may yield even more pronounced differences between pathway resources. Additionally, further techniques for gene expression-based pathway activity scoring could be incorporated, such as Pathifier or SAS (Drier et al., 2013; Lim et al., 2016).

DATA AVAILABILITY STATEMENT

All datasets generated/analyzed for this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

DD-F conceived and designed the study. SM and DD-F conducted the main analysis and implemented the Python package. HF supervised methodological aspects of the analysis. CH and AG assisted technically in the analysis of the results. MH-A acquired the funding. SM, HF, CH, MH-A, and DD-F wrote the paper.

FUNDING

This work was supported by the EU/EFPIA Innovative Medicines Initiative Joint Undertaking under AETIONOMY (grant number 115568), resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies in kind contribution.

ACKNOWLEDGMENTS

The authors would like to thank Mohammad Asif Emon for his assistance in conducting SPIA and Jan-Eric Bökenkamp for his assistance in processing the TCGA datasets. Furthermore, we would like to thank Jonas Klees and Carina Steinborn for generating the visuals in this paper. Finally, we would like to thank the curators of KEGG, Reactome, and WikiPathways as well as the TCGA network for generating the pathway content and datasets used in this work, respectively.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2019.01203/full#supplementary-material>

REFERENCES

- Bader, G. D., Cary, M. P., and Sander, C. (2006). Pathguide: a pathway resource list. *Nucleic Acids Res.* 34 (suppl_1), D504–D506. doi: 10.1093/nar/gkj126
- Barbie, D. A., Tamayo, P., Boehm, J. S., Kim, S. Y., Moody, S. E., Dunn, I. F., et al. (2009). Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* 462 (7269), 108. doi: 10.1038/nature08460
- Bateman, A. R., El-Hachem, N., Beck, A. H., Aerts, H. J., and Haibe-Kains, B. (2014). Importance of collection in gene set enrichment analysis of drug response in cancer cell lines. *Sci. Rep.* 4, 4092. doi: 10.1038/srep04092
- Bayerlová, M., Jung, K., Kramer, F., Klemm, F., Bleckmann, A., and Beißbarth, T. (2015). Comparative study on gene set and pathway topology-based enrichment methods. *BMC Bioinf.* 16 (1), 334. doi: 10.1186/s12859-015-0751-5
- Belinky, F., Nativ, N., Stelzer, G., Zimmerman, S., Iny Stein, T., Safran, M., et al. (2015). PathCards: multi-source consolidation of human biological pathways. *Database* 2015. doi: 10.1093/database/bav006
- Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B. (Methodological)* 57 (1), 289–300. doi: 10.2307/2346101
- Benjamini, Y., and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* 29 (4), 1165–1188. doi: 10.1214/aos/1013699998
- Cancer Genome Atlas Research Network. (2014). Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* 513 (7517), 202. doi: 10.1038/nature13480
- Cerami, E. G., Gross, B. E., Demir, E., Rodchenkov, I., Babur, O., Anwar, N., et al. (2011). Pathway commons, a web resource for biological pathway data. *Nucleic Acids Res.* 39 (Suppl. 1), D685–D690. doi: 10.1093/nar/gkq1039
- Coates, A. S., Winer, E. P., Goldhirsch, A., Gelber, R. D., Gnant, M., Piccart-Gebhart, M., et al. (2015). Tailoring therapies—improving the management of early breast cancer: St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2015. *Ann. Oncol.* 26 (8), 1533–1546. doi: 10.1093/annonc/mdv221
- Colaprico, A., Silva, T. C., Olsen, C., Garofano, L., Cava, C., Garolini, D., et al. (2015). TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* 44 (8), e71–e71. doi: 10.1093/nar/gkv1507
- Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* 20 (3), 273–297. doi: 10.1007%2FBF00994018
- Doderer, M. S., Anguiano, Z., Suresh, U., Dashnamoorthy, R., Bishop, A. J., and Chen, Y. (2012). Pathway Distiller-multisource biological pathway consolidation. *BMC Genom.* 13 (6), S18. doi: 10.1186/1471-2164-13-S6-S18
- Domingo-Fernández, D., Hoyt, C. T., Bobis-Álvarez, C., Marin-Llao, J., and Hofmann-Apitius, M. (2018). ComPath: an ecosystem for exploring, analyzing, and curating mappings across pathway databases. *NPJ Syst. Biol. Appl.* 4 (1), 43. doi: 10.1038/s41540-018-0078-8
- Domingo-Fernandez, D., Mubeen, S., Marin-Llao, J., Hoyt, C., and Hofmann-Apitius, M. (2019). PathMe: merging and exploring mechanistic pathway knowledge. *BMC Bioinf.* 20, 243. doi: 10.1186/s12859-019-2863-9
- Drier, Y., Sheffer, M., and Domany, E. (2013). Pathway-based personalized analysis of cancer. *Proc. Nat. Acad. Sci.* 110 (16), 6388–6393. doi: 10.1073/pnas.1219651110
- Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., et al. (2018). The reactome pathway knowledgebase. *Nucleic Acids Res.* 46 (D1), D649–D655. doi: 10.1093/nar/gkx1132
- Fabris, F., Palmer, D., de Magalhães, J. P., and Freitas, A. A. (2019). Comparing enrichment analysis and machine learning for identifying gene properties that discriminate between gene classes. *Briefings Bioinf.* doi: 10.1093/bib/bbz028
- Fisher, R. A. (1992). Statistical methods for research workers in *Breakthroughs in Statistics* (New York, NY:Springer), 66–70.
- Fröhlich, H. (2014). Including network knowledge into Cox regression models for biomarker signature discovery. *Biom. J.* 56 (2), 287–306. doi: 10.1002/bimj.201300035
- Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Software* 33 (1), 1. doi: 10.18637/jss.v033.i01
- García-Campos, M. A., Espinal-Enríquez, J., and Hernández-Lemus, E. (2015). Pathway analysis: state of the art. *Front. Physiol.* 6, 383. doi: 10.3389/fphys.2015.00383
- Grüning, B. A., Lampa, S., Vaudel, M., and Blankenberg, D. (2019). Software engineering for scientific big data analysis. *GigaScience* 8 (5), giz054. doi: 10.1093/gigascience/giz054
- Graudenzi, A., et al. (2017). Pathway-based classification of breast cancer subtypes. *Front. Biosci., (Landmark Ed)* 22, 1697–1712. doi: 10.2741/4566
- Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., and Rosati, R. A. (1982). Evaluating the yield of medical tests. *JAMA* 247 (18), 2543–2546. doi: 10.1001/jama.1982.03320430047030
- Hoyt, C. T., Konotopez, A., and Ebeling, C. (2018). PyBEL: a computational framework for Biological Expression Language. *Bioinformatics* 34 (4), 703–704. doi: 10.1093/bioinformatics/btx660
- Hoyt, C. T., Domingo-Fernández, D., Mubeen, S., Llaó, J. M., Konotopez, A., Ebeling, C., et al. (2019). Integration of Structured Biological Data Sources using Biological Expression Language. *Biorxiv* 631812. doi: 10.1101/631812
- Ihnatova, I., Popovici, V., and Budinska, E. (2018). A critical comparison of topology-based pathway analysis methods. *PLoS One* 13 (1), e0191154. doi: 10.1371/journal.pone.0191154
- Kamburov, A., Wierling, C., Lehrach, H., and Herwig, R. (2008). ConsensusPathDB—a database for integrating human functional interaction networks. *Nucleic Acids Res.* 37 (suppl_1), D623–D628. doi: 10.1093/nar/gkn698
- Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2016). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 45 (D1), D353–D361. doi: 10.1093/nar/gkw1092
- Khatri, P., Sirota, M., and Butte, A. J. (2012). Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.* 8 (2), e1002375. doi: 10.1371/journal.pcbi.1002375
- Kirouac, D. C., Saez-Rodriguez, J., Swantek, J., Burke, J. M., Lauffenburger, D. A., and Sorger, P. K. (2012). Creating and analyzing pathway and protein interaction compendia for modelling signal transduction networks. *BMC Syst. Biol.* 6 (1), 29. doi: 10.1186/1752-0509-6-29
- Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. (2015). The molecular signatures database hallmark gene set collection. *Cell Syst.* 1 (6), 417–425. doi: 10.1016/j.cels.2015.12.004
- Lim, S., Lee, S., Jung, I., Rhee, S., and Kim, S. (2018). Comprehensive and critical evaluation of individualized pathway activity measurement tools on pan-cancer data. *Briefings Bioinf.*
- Lim, S., Park, Y., Hur, B., Kim, M., Han, W., and Kim, S. (2016). Protein interaction network (pin)-based breast cancer subsystem identification and activation measurement for prognostic modeling. *Methods* 110, 81–89. doi: 10.1016/j.ymeth.2016.06.015
- Mayr, A., and Schmid, M. (2014). Boosting the concordance index for survival data—a unified framework to derive and evaluate biomarker combinations. *PLoS One* 9 (1), e84483. doi: 10.1371/journal.pone.0084483
- McKinney, W. (2010). Data Structures for Statistical Computing in Python in *Proceedings of the 9th Python in Science Conference*. Eds. van der Walt, S., and Millman, J., 51–56.
- Miller, R. A., Ehrhart, F., Eijssen, L. M., Slenter, D. N., Curfs, L. M., Evelo, C. T., et al. (2019). Beyond pathway analysis: Identification of active subnetworks in Rett syndrome. *Front. Genet.* 10, 59. doi: 10.3389/fgene.2019.00059
- Molinari, A. M., Simon, R., and Pfeiffer, R. M. (2005). Prediction error estimation: a comparison of resampling methods. *Bioinformatics* 21 (15), 3301–3307. doi: 10.1093/bioinformatics/bti499
- Povey, S., Lovering, R., Bruford, E., Wright, M., Lush, M., and Wain, H. (2001). The HUGO gene nomenclature committee (HGNC). *Hum. Genet.* 109 (6), 678–680. doi: 10.1007/s00439-001-0615-0
- Reimand, J., Isserlin, R., Voisin, V., Kucera, M., Tannus-Lopes, C., Rostamianfar, A., et al. (2019). Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nat. Protoc.* 14 (2), 482–517. doi: 10.1038/s41596-018-0103-9
- Sales, G., Calura, E., and Romualdi, C. (2018). meta Graphite—a new layer of pathway annotation to get metabolite networks. *Bioinformatics* 35 (7), 1258–1260. doi: 10.1093/bioinformatics/bty719

- Schaefer, C. F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., et al. (2008). PID: the pathway interaction database. *Nucleic Acids Res.* 37 (suppl_1), D674–D679. doi: 10.1093/nar/gkn653
- Senkus, E., Kyriakides, S., Ohno, S., Penault-Llorca, F., Poortmans, P., Rutgers, E., et al. (2015). Primary breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.* 26 (suppl_5), v8–v30. doi: 10.1093/annonc/mdv298
- Slater, T. (2014). Recent advances in modeling languages for pathway maps and computable biological networks. *Drug Discovery Today* 19 (2), 193–198. doi: 10.1016/j.drudis.2013.12.011
- Slater, D. N., Kutmon, M., Hanspers, K., Riutta, A., Windsor, J., Nunes, N., et al. (2017). WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res.* 46 (D1), D661–D667. doi: 10.1093/nar/gkx1064
- Sorlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci.* 98 (19), 10869–10874. doi: 10.1073/pnas.191367098
- Stoney, R. A., Schwartz, J. M., Robertson, D. L., and Nenadic, G. (2018). Using set theory to reduce redundancy in pathway sets. *BMC Bioinf.* 19 (1), 386. doi: 10.1186/s12859-018-2355-3
- Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Nat. Acad. Sci.* 102 (43), 15545–15550. doi: 10.1073/pnas.0506580102
- Tarca, A. L., Draghici, S., Khatri, P., Hassan, S. S., Mittal, P., Kim, J. S., et al. (2008). A novel signaling pathway impact analysis. *Bioinformatics* 25 (1), 75–82. doi: 10.1093/bioinformatics/btn577
- Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Stat. Med.* 16 (4), 385–395. doi: 10.1002/(sici)1097-0258(19970228)16:4<385::aid-sim380>3.0.co;2-3
- Türei, D., Korcsmáros, T., and Saez-Rodriguez, J. (2016). OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat. Methods* 13 (12), 966. doi:10.1038/nmeth.4077
- Van Wieringen, W. N., Kun, D., Hampel, R., and Boulesteix, A. L. (2009). Survival prediction using gene expression data: a review and comparison. *Comput. Stat. Data Anal.* 53 (5), 1590–1603. doi: 10.1016/j.csda.2008.05.021
- Vivar, J. C., Pemu, P., McPherson, R., and Ghosh, S. (2013). Redundancy control in pathway databases (ReCiPa): an application for improving gene-set enrichment analysis in Omics studies and "Big data" biology. *Omics: J. Integr. Biol.* 17 (8), 414–422. doi: 10.1089/omi.2012.0083
- Wadi, L., Meyer, M., Weiser, J., Stein, L. D., and Reimand, J. (2016). Impact of outdated gene annotations on pathway enrichment analysis. *Nat. Methods* 13 (9), 705. doi: 10.1038/nmeth.3963
- Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* 45 (10), 1113. doi: 10.1038/ng.2764
- Zhang, Y., Yang, W., Li, D., Yang, J. Y., Guan, R., and Yang, M. Q. (2018). Toward the precision breast cancer survival prediction utilizing combined whole genome-wide expression and somatic mutation analysis. *BMC Med. Genom.* 11 (5), 104. doi: 10.1109/BIBM.2017.8217762
- Zou, H., and Trevor, H. (2005). Regularization and Variable Selection via the Elastic Net. *J. R. Stat. Soc. Ser. B: 67* (2), 301–320. doi: 10.1111/j.1467-9868.2005.00503.x

Conflict of Interest: HF received salaries from UCB Biosciences GmbH. UCB Biosciences GmbH had no influence on the content of this work.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.






Copyright © 2019 Mubeen, Hoyt, Gemünd, Hofmann-Apitius, Fröhlich and Domingo-Fernández. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

A.3 DecoPath: a web application for decoding pathway enrichment analysis

Reprinted with permission from “Mubeen S., Bharadhwaj S. V., Kodamullil A.T., Gadiya Y., Hofmann-Apitius M., and Domingo-Fernández D. (2021). DecoPath: A Web Application for Decoding Pathway Enrichment Analysis. *NAR Genomics and Bioinformatics*, 3(3): lqab087”.

Copyright © Mubeen, S., *et al.*, 2021.

DecoPath: a web application for decoding pathway enrichment analysis

Sarah Mubeen ^{1,2,3,*}, Vinay S. Bharadhwaj ^{1,2}, Yojana Gadiya ^{1,2},
Martin Hofmann-Apitius ^{1,2}, Alpha T. Kodamullil ¹ and Daniel Domingo-Fernández ^{1,3,4,*}

¹Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing, Sankt Augustin 53757, Germany, ²Bonn-Aachen International Center for IT, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn 53115, Germany, ³Fraunhofer Center for Machine Learning, Germany and ⁴Enveda Biosciences, Boulder, CO 80301, USA

Received June 10, 2021; Revised August 31, 2021; Editorial Decision September 07, 2021; Accepted September 14, 2021

ABSTRACT

The past decades have brought a steady growth of pathway databases and enrichment methods. However, the advent of pathway data has not been accompanied by an improvement in interoperability across databases, hampering the use of pathway knowledge from multiple databases for enrichment analysis. While integrative databases have attempted to address this issue, they often do not account for redundant information across resources. Furthermore, the majority of studies that employ pathway enrichment analysis still rely upon a single database or enrichment method, though the use of another could yield differing results. These shortcomings call for approaches that investigate the differences and agreements across databases and methods as their selection in the design of a pathway analysis can be a crucial step in ensuring the results of such an analysis are meaningful. Here we present DecoPath, a web application to assist in the interpretation of the results of pathway enrichment analysis. DecoPath provides an ecosystem to run enrichment analysis or directly upload results and facilitate the interpretation of results with custom visualizations that highlight the consensus and/or discrepancies at the pathway- and gene-levels. DecoPath is available at <https://decopath.scai.fraunhofer.de>, and its source code and documentation can be found on GitHub at <https://github.com/DecoPath/DecoPath>.

INTRODUCTION

In recent years, high-throughput (HT) technologies have given rise to a perpetual influx of *-omics* data, requiring pragmatic approaches to sift out meaning. One of the most

common applications of HT technologies is gene expression profiling to simultaneously determine the expression patterns of thousands of genes at the transcription level under certain conditions (1). While a host of statistical techniques are available to identify genes that differ in expression depending on a particular condition, gene set or pathway enrichment analysis methods represent a major class of tools researchers employ to group lists of genes into defined pathways and understand the functional roles of genes for any given set of conditions (2). To date, almost a hundred different pathway enrichment methods have been proposed, including the popular over-representation analysis (ORA) and gene set enrichment analysis (GSEA) (3). Though these methods may vary based on the overarching categories they fall into (e.g. topology versus non-topology-based) or the statistical techniques used, they have widely shown their ability to deconvolute biological pathways dysregulated in a given state (4).

Numerous pathway databases have been developed which aim at representing biological pathways from various vantage points (e.g. differing scopes, contexts, boundaries or pathway types). The existence of several hundreds of these databases reflects the inherent complexity and variability of biological processes that occur in living organisms (5). Further compounding this complexity is the fact that biological pathways housed in these databases are human constructs, delimited based on abstract boundaries defined by a researcher or the consensus of the community. This implies that a well-studied pathway could contain different biological entities depending on the boundaries defined by the databases that store it. These differences across databases can manifest in variability in the results of pathway enrichment analysis (6,7), in a similar way as methods can impact results (4,8–10).

Recent approaches to pathway enrichment analysis have focused on the integration of multiple datasets across different platforms to ensure a broader coverage of significantly enriched pathways (11–13). Other techniques attempt to

*To whom correspondence should be addressed. Tel: +49 2241 14 4036; Email: sarah.mubeen@scai.fraunhofer.de
Correspondence may also be addressed to D. Domingo-Fernández. Email: daniel.domingo.fernandez@scai.fraunhofer.de

account for potential differences that may arise in the results of pathway enrichment analysis by combining gene sets from several pathway databases. For instance, (14) presented an approach that leverages GSEA to calculate a combined enrichment score for multiple *-omics* layers using several databases. However, performing pathway enrichment analysis using multiple databases to increase the number of pathways covered can only partially address the challenges associated with variability in results. This is because such an approach falls short of leveraging the substantial overlap of pathway knowledge across databases which could provide more comprehensive results (15–17) or shed light on inconsistencies across pathway databases (18). Furthermore, combining several databases can result in redundant pathways, an issue tackled by the SetRank algorithm which discounts significant gene sets if their significance can be explained by their overlap with another gene set (19). Finally, a possible, natural solution to better connect and structure redundant information across databases lies in leveraging pathway ontologies (20) or pathway mappings with database cross-references (17). By connecting related pathways across databases, we can, in turn, investigate the consensus, or lack thereof, of the results of pathway enrichment analysis between databases or methods as demonstrated by several recent benchmarks (4,8–10).

Here, we present DecoPath, a web application that provides a user-friendly and interactive application to compare and interpret the results of pathway enrichment analysis yielded by different pathway databases. To facilitate the comparison of results across databases and bring to light possible contradictory results, we present several interactive visualization tools designed to better interpret the results of pathway enrichment at both the pathway and gene-level. While these visualizations can generally be used for any pathway enrichment method, DecoPath also integrates standard pathway enrichment methods in its pipeline, thus, enabling users to conduct an entire enrichment analysis on the web application (from data submission to interpretation). Finally, although DecoPath provides four default databases, it also allows users to upload gene sets and mappings such that analyses can be run on their independently curated gene sets.

MATERIALS AND METHODS

Implementation

The server-side was implemented in the Python programming language using the Django framework (<https://www.djangoproject.com/>). This framework operates using a Model-View-Controller (MVC) architecture and was integrated with Celery (<http://www.celeryproject.org>) and RabbitMQ (<https://www.rabbitmq.com>) for asynchronous task execution. The front-end of DecoPath comprises several interactive visualizations implemented using a collection of powerful Javascript libraries, including jQuery (<https://jquery.com>), D3.js (<https://d3js.org/>) and DataTables (<https://datatables.net/>). Furthermore, DecoPath relies on Bootstrap 4 (<https://getbootstrap.com/>) for the main design of the website. The web application is containerized using Docker for reproducibility purposes and easy deployment. We strongly recommend the use of DecoPath on Chrome,

Firefox or Safari browsers and on Mac or Linux operating systems.

Pathway resources

DecoPath enables users to compare the results of enrichment analysis yielded using various pathway databases. As mentioned in the Introduction, pathways in different databases can substantially overlap, such that a pathway in one database can have counterparts in several others. Leveraging equivalent pathway mappings across several widely-used databases, DecoPath aims at highlighting the consensus, or lack thereof, of enrichment analysis results for each equivalent pathway. Expanding upon our previous work (17), we added novel equivalent pathway mappings as well as mappings for an additional database (i.e. PathBank (21)) (Supplementary Text). Thus, the released version of DecoPath provides users with the following pathway databases: KEGG (22), Reactome (23), WikiPathways (24) and PathBank (Retrieved 3 August 2020). Additionally, as integrative resources can lead to more biologically consistent results in enrichment analysis (6), a DecoPath-specific gene set database containing merged gene sets of equivalent pathways across the aforementioned databases is also provided, as described in the following section. Finally, in order to ensure that regular updates to these pathway resources are reflected in DecoPath, the software is updated with the latest gene sets annually.

Generating a pathway hierarchy

The consolidation of each of the pathway databases into a pathway meta-database was conducted in order to generate a pathway hierarchy. In doing so, equivalent representations of pathways across KEGG, PathBank, Reactome and WikiPathways were combined. The pathway hierarchy contains a total of 644 pathways from these four databases and can be found at https://github.com/ComPath/compath-resources/blob/master/mappings/decopath_ontology.xlsx (dated 13 January 2021). The hierarchy comprises eight major categories: metabolism, immune, signaling, communication and transport, cell death, disease, DNA repair and replication, and others. All pathways in the hierarchy retained their original identifiers except equivalent pathways which were merged and given unique names and identifiers. The pathway hierarchy is a directed acyclic graph with a maximum depth of 4, in which relation types between pathways can be either *is-part-of* or *equivalent-to* relations. The curation process to generate the hierarchy is described in the Supplementary Text. Periodic updates to the pathway hierarchy are made on an annual basis.

Pathway enrichment methods

DecoPath comprises two of the most widely used pathway enrichment methods (25–27): over representation analysis (ORA) and gene set enrichment analysis (GSEA) (3). ORA aims at identifying pathways (i.e. gene sets) that are over-represented within a list of genes of interest. A pathway is considered enriched (over-represented) if the *P*-value arising from a one-sided Fisher's exact test (28)

is lower than a specified threshold, typically 0.05. As this test is conducted for each pathway in the database, DecoPath's implementation of ORA corrects the P -value by applying multiple hypothesis testing correction with the Benjamini–Yekutieli method under dependency (29). The second method, GSEA, determines whether a pathway or a gene set significantly differs between two groups. A pathway is considered significantly regulated in that condition if genes of that pathway appear in the top or bottom ranking of a list of differentially expressed genes (DEGs) more than expected by chance. An alternative version of GSEA, namely GSEA Pre-Ranked (3), is also available if users wish to run GSEA on a pre-ranked list of genes. DecoPath uses implementations of GSEA and GSEA Pre-Ranked from *gseapy* (<https://gseapy.readthedocs.io/en/latest>). Additionally, DecoPath enables conducting differential gene expression (DGE) analysis between groups through DESeq2 (version 1.22.2). Apart from these methods, DecoPath also provides the option to include additional pathway enrichment methods into the web application.

Installation

Although we provide a freely available instance of DecoPath at <https://decopath.scai.fraunhofer.de/>, in the case of large datasets or cases where the compute capacity of the server may be insufficient depending on the type of analysis, users can install and use DecoPath in their own system. We offer two options to install DecoPath depending on the needs of the user. The first and easiest method for those unfamiliar with Django-based web applications is to install Docker and deploy the Docker container which will install required components and run the web application. Detailed instructions are provided on GitHub (<https://github.com/decopath/decopath>). Alternatively, DecoPath can be directly deployed following the instructions in the GitHub repository.

Runtime considerations

Computation time is dependent on the type of analysis, size of the datasets as well as the device specifications. ORA can be run on a gene list on a timescale of seconds and requires the relatively lowest usage of memory. A DGE analysis task has a timescale of several minutes, while GSEA on a typical expression dataset with two experimental groups and four databases can also be done within minutes with a dual-core Intel Core i5 CPU and 16 GB RAM.

Case scenario

Using each of the available enrichment methods, we demonstrate a typical workflow in DecoPath with the The Cancer Genome Atlas Liver Hepatocellular Carcinoma (TCGA-LIHC) dataset (30). Gene expression data from this dataset was retrieved from the Genomic Data Commons (GDC; <https://gdc.cancer.gov>) portal through the R/Bioconductor package, TCGAbiolinks (version 2.16.3; (31)) on 4 August 2020. To run GSEA, we employed RNA-Seq expression data normalized using Fragments Per Kilobase of transcript per Million mapped reads upper quartile (FPKM-

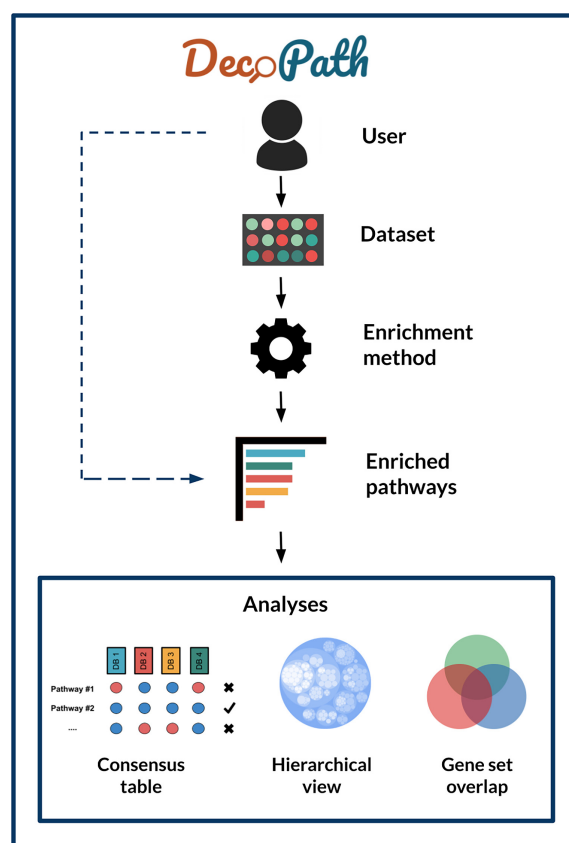


Figure 1. DecoPath workflow. Users can upload datasets to run pathway enrichment analysis or directly upload enrichment results from their own experiments. Once results have been loaded, DecoPath offers users several visualizations designed to evaluate pathway consensus at the database, hierarchy and gene set level. Users can also opt to directly upload results generated from varying enrichment methods across to visualize variations from these against a set of pathway databases.

UQ). DGE analysis using read counts from the TCGA-LIHC dataset (retrieved from the GDC; <https://gdc.cancer.gov>) was performed between normal and tumor samples to derive a gene list to conduct ORA. This final list of genes was restricted to genes that exhibited an adjusted P -value < 0.05 . Specifications of the parameter settings for ORA and GSEA are listed in Supplementary Table S1.

RESULTS

Here, we describe the DecoPath web application. A typical workflow of the web application involves the submission of an experiment, generation of results, and the subsequent exploration and visualization of these results (Figure 1). In the following, we provide a detailed description for each of the steps in the workflow.

Submission form

Once a user has logged into DecoPath, on the Homepage, the input form allows them to upload their files and select parameters to run different analyses or upload results from them (Figure 2). For users opting to run analyses using De-



Figure 2. DecoPath homepage. Once a user has logged in, on the homepage, they are provided with the option to either run or submit the results of a pathway analysis. If a user opts to submit the results of an analysis, they can upload their data, select the databases they wish to include, choose the parameter settings for each experiment and optionally perform a concurrent DGE analysis. Once the form has been submitted, users are directed to the Experiments page where they can find visualizations and functionalities to compare and explore the consensus around different pathway databases.

coPath, the workflow depends on the analysis they select. Briefly, GSEA requires the submission of datasets, such as from RNA-Seq, microarray or ChIP-Seq, accompanied by a design matrix denoting the class labels (e.g. normal and tumor) for samples in the dataset. To run ORA, users need only submit a list of genes of interest. For either method, users can select which of the four pathway databases they would like to include in the analysis. By default, genesets from DecoPath which contain merged equivalent pathways are also included in the analysis.

These pathway enrichment methods can also be supplemented by DGE analysis to generate visualizations and identify genes that are differentially expressed according to a fold change cutoff. In order to run DGE analysis, unnormalized read counts in the form of a matrix of integer values is required, as is a design matrix, analogous to the one required for GSEA. For each of these analyses, gene identifiers should be in the form of HUGO Gene Nomenclature Committee (HGNC) symbols. Alternatively, users can opt to download gene set files for pathway databases included in DecoPath, run GSEA, ORA and/or DGE analysis, and upload the results of the analysis to the website. By directly uploading the results, users can also analyze the results of alternative enrichment methods such as Enrich-Net (32) and Signaling Pathway Impact Analysis (SPIA) (33) using DecoPath. Detailed descriptions of the input files can be found in the User Guide and FAQs sections on our website.

Visualizations and analyses

Once users have submitted their query, they are directed to the Experiments page where they can view the status as well as details of their experiments, and explore and visualize their results (Figure 3). To interpret the results of enrichment analysis, we implemented multiple, customized tools intended to provide insights on the consensus across databases, each of which we detail below.

Exploring the consensus across pathway databases

The first visualization summarizes the consensus results of pathway enrichment analysis on multiple databases. For each pathway (row), the table shows the concordance across databases, reflected in terms of the significance value, specifically for ORA, and both the significance value and directionality of the normalized enrichment score (NES) for GSEA (Figure 4). Using this visualization, users can rapidly identify concordant (i.e. a given pathway is reported as significantly enriched in a gene list across all databases) and contradictory (i.e. a given pathway is reported as significantly enriched in a gene list in one or more databases, but not in the others [or vice versa]) pathways and directly compare their results.

We conducted a case scenario to investigate the results for ORA and GSEA using four pathway databases on the TCGA-LIHC dataset. Among the pathways enriched in ORA which could be found in more than one pathway database, we found 88 concordant pathways and 41 contradictory ones. Similarly, the results of GSEA revealed 70 concordant and 45 contradictory pathways. Among the contradictory pathways we observed in GSEA, the majority of contradictions pertained to whether or not the pathway was significantly enriched, while 12 pathways also differed in the sign of the NES (i.e. the same pathway was reported as enriched at the top of a ranked gene list for one database and at the bottom for another). Additionally, 53 concordant pathways were common between the results of GSEA and ORA; however, as expected, differences based on the pathway enrichment method were observed. Overall, the results of the LIHC-TCGA dataset for both methods showed that approximately one-third of equivalent pathways were contradictory across the two methods. Thus, the selection of databases, as well as the enrichment method, are important aspects in the experimental design of pathway enrichment analysis. We have observed that the use of one over another can yield discordant results, leading to different interpretations of results depending on the database choice. In the following sections, we illustrate why these results may be discrepant by analyzing the gene sets of a given pathway.

Visualizing consensus through the pathway hierarchy

In the second visualization, users can explore the results of their analysis within the context of a pathway hierarchy (see Materials and Methods section). This user-friendly and interactive visualization represents the different levels of the pathway hierarchy as circles, each of which represent a child or a parent pathway. In the case of GSEA, pathways that do not show statistically significant (adjusted P -value < 0.05)

Home User Guide FAQs Experiments Account Logout

DecoPath Experiments

Here, you can view details of your experiments, explore results and cancel or delete experiments.

The status of the experiment is indicated in the "Status" column as follows:

- ✓ The experiment has run successfully. Results are ready for viewing, exploring and downloading.
- 🔄 The experiment is in progress.
- ✗ The experiment has failed.

Experiment	Status	Results	Consensus	Explore	Analysis	Databases	Dataset	Classes
1	✓	Load Results	Load Consensus Table	Visualize Consensus	GSEA	DecoPath KEGG PathBank Reactome WikiPathways	final_results.tsv	NA
2	✓	Load Results	Load Consensus Table	Visualize Consensus	ORA	DecoPath KEGG PathBank Reactome WikiPathways	lihc_read_count_duplicates_removed.csv	tumor normal
3	🔄	Load Results	Load Consensus Table	Visualize Consensus	GSEA	DecoPath KEGG PathBank Reactome WikiPathways	lung_expression.tsv	Tumor Normal
4	✗	Load Results	Load Consensus Table	Visualize Consensus	GSEA	DecoPath KEGG Reactome WikiPathways	heart_expression.tsv	Tumor Normal

Figure 3. Experiments page. The Experiments page lists details of each of the experiments that were run or uploaded. The status of the experiment is given in the 'Status' column, indicating whether the experiment was successfully run, if it is pending or has failed. Through this page, users can then navigate to each of the different visualizations to explore the results of their analysis.

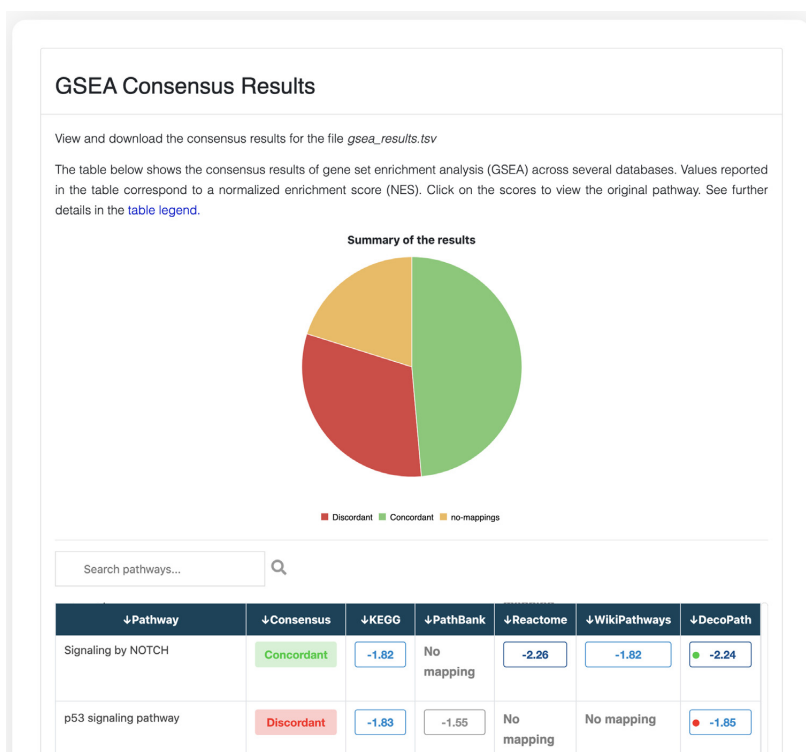


Figure 4. Consensus page. The Consensus page visualization shows the consensus of the results of enrichment analysis across databases at the pathway level. In the case of GSEA, the table displays the NES for a given pathway across each database as well as the NES of the merged gene sets of all equivalent pathways, the latter of which is indicated in the column 'DecoPath'.

differences between groups are colored gray, while statistically significant ones are colored red or blue based on the sign of the NES, and shaded by a gradient based on the magnitude of the NES. In the case of ORA, pathways are colored gray if they are not significant with an adjusted P -value < 0.05 and red otherwise. Additionally, the size of the gene sets for each of the pathways is proportional to the size of the circles. Furthermore, interactive visualizations also offer zoom and search functionalities to easily identify pathways of interest. In summary, with this tool, users can not only explore the enrichment results through the entire pathway hierarchy but also intuitively evaluate equivalent pathways and the size of the pathways, both of which are known to affect results (6,34).

Continuing the case scenario on the LIHC datasets, this visualization was used to identify major pathways that were enriched in both ORA and GSEA (Figure 5). The organization of pathways into eight major categories allows users to intuitively navigate through the hierarchy and identify pathway groups in which several pathways are enriched. For instance, among all pathways pertaining to metabolism, we observed that lipid and purine metabolism pathways were significantly enriched in both GSEA and ORA, indicating that there was a consensus across both methods and databases. Among other examples of consensus, we found cytokine signaling within the immune system pathways as well as MAP kinase signaling within the signaling pathways significantly enriched in all methods and databases. Finally, contrasting colors of this hierarchical view allow for the rapid identification of contradictory pathways which can then be further analyzed at the gene-level, aided by the following visualization.

Analyzing equivalent pathways at the gene level

The third visualization is an interactive Venn diagram that shows the overlap for equivalent pathways at the gene-level. In this visualization, we provide a means to analyze exactly which genes may explicate the findings of the pathway analysis. By clicking on the subsets of the Venn diagram, users can display the genes in each of the gene sets. Thus, users can pinpoint the specific genes of the pathway that might contribute to the contradictions observed in the results of the enrichment analysis. If fold changes have additionally been uploaded of DEGs or DGE analysis has been performed, users can also view the distribution of fold changes of genes in the dataset in an accompanying histogram.

To demonstrate this visualization, we explored both a pathway showing concordant results (i.e. DNA replication pathway) and another showing contradictory results (pyruvate metabolism) from the results of pathway enrichment on the TCGA-LIHC dataset. In the case of the DNA replication pathway, the results showed that the KEGG, Reactome and WikiPathways equivalent representations consistently reported NES over 2.0, suggesting that the pathway is regulated in the liver cancer dataset. We then explored the overlap of the gene sets of the DNA replication pathway from the three databases, observing that the \log_2 fold change values for the vast majority of genes in the pathway were positive. As GSEA finds the pathways which are nearest to the top (or bottom) of the ranked list of DEGs, this can account

for the observance of the high NES (Figure 6A). Similarly, we explored a pathway (i.e. pyruvate metabolism), which had contradictory results in KEGG, Reactome and PathBank. In this case, these pathway databases disagreed in the direction of regulation of the NES; while the NES of pyruvate metabolism was positive in KEGG and PathBank, the sign of the NES was negative in Reactome. The consensus between KEGG and PathBank is not surprising as the gene sets of the pathway largely overlap (Figure 6B), while only 13 of the 31 genes in the Reactome pathway overlap with the other two gene sets. By plotting the distribution of the other 18 genes that are uniquely present in the Reactome pathway, we found that these genes were largely over-expressed, explaining the observed differences in the NES between them. Thus, this example illustrates how this tool can be used to assist in the interpretation of the discrepant results of pathway enrichment analysis.

DISCUSSION

While the popularity of pathway enrichment analysis for the interpretation of *-omics* data has grown over the past two decades and led to the development of over a hundred different methods, recent benchmarks have shown that the selected method can influence results (4,8,9,27). Furthermore, the majority of pathway enrichment analyses tend to be conducted on a single pathway database, the choice of which can also impact results of an analysis (6). While several tools have been implemented to run enrichment analysis on multiple platforms and methods (see Introduction), tools that facilitate the direct comparison of results yielded using different databases or enrichment methods at the pathway- and gene-levels are lacking. To address this issue, we have presented DecoPath, the first web application designed to assist in the interpretation of the results of pathway enrichment methods. DecoPath provides users with a broad range of built-in tools and visualization to conduct enrichment analyses and guide them in the interpretation of the results using multiple pathway databases.

Nonetheless, the presented web application is not without its limitations. First, while multiple enrichment methods exist, DecoPath only enables running two of the most popular pathway enrichment analyses. Similarly, DecoPath exclusively contains four pathway databases given the substantial curation effort required to map and harmonize pathway databases. To address these limitations, we enable users to directly upload results from other enrichment methods or pathway mappings from additional databases. Another limitation is the computational power of the server required to run experiments on datasets with a large sample size, or depending on the type of analysis conducted, may not be enough. However, since the source code of the web application is available (<https://github.com/DecoPath/DecoPath>) and DecoPath can be containerized in Docker, users can deploy the web application as per their needs to run more computationally demanding analyses.

In the future, we plan to map and integrate additional databases into DecoPath, as well as more enrichment methods. Furthermore, we envision the implementation of a consensus algorithm to combine the results obtained across multiple databases into a single score, in line with ap-

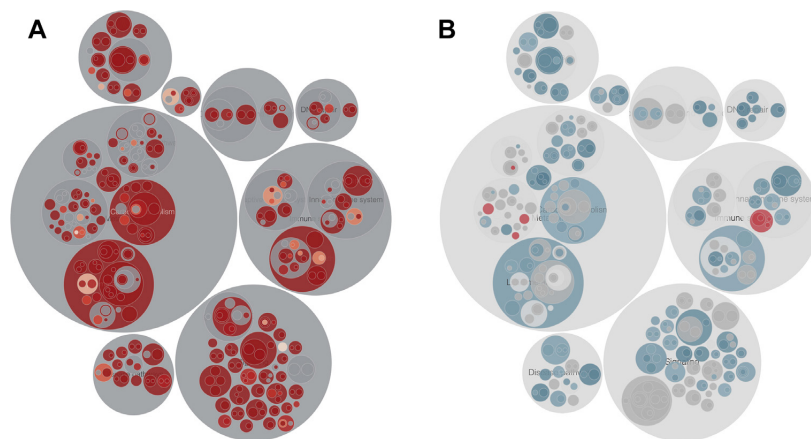


Figure 5. Circle pack visualization of the pathway hierarchy using different pathway enrichment methods. The figure corresponds to the interactive visualizations displaying the results of running ORA (A) and GSEA (B) on the LIHC dataset. In this visualization, results are customized based on the pathway enrichment method. In the case of Functional Class Scoring (FCS) and Pathway Topology (PT) based methods, the visualization highlights the direction of the dysregulation for each significantly dysregulated pathway as well as for the adjusted *P*-value (B). On the other hand, for ORA, the visualization highlights pathways that are significantly enriched based on an adjusted *P*-value (A).

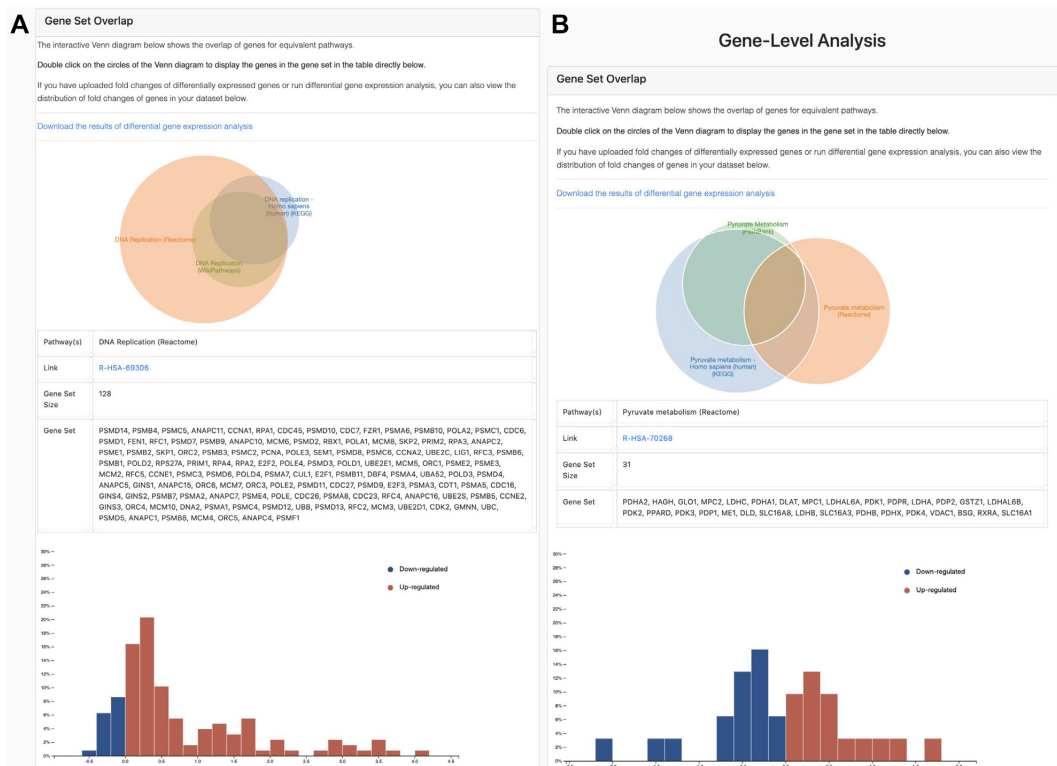


Figure 6. Overlap of gene sets for a given pathway. Venn diagrams display the overlap of gene sets for equivalent pathways across user selected databases. By running DGE analysis, users can also view a histogram of the distribution of \log_2 fold changes for DEGs in their dataset to identify which genes are leading to either consistent or contradictory results for their pathway analysis. (A) Venn diagram of the overlap of gene sets for the DNA replication pathway from KEGG, Reactome and WikiPathways is shown above, while a histogram of \log_2 fold changes for DEGs from this pathway is shown below (in this example, the pathway representation from Reactome). (B) Venn diagram of the pyruvate metabolism pathway from KEGG, Reactome and PathBank and a histogram of \log_2 fold changes for DEGs for the pyruvate metabolism pathway Reactome are displayed.

proaches which integrate results obtained by an ensemble of enrichment methods, such as CGPS (35) and EGSEA (36), whilst taking into account variables such as gene set size and the magnitude of the enrichment score and/or *P*-value. Finally, we hope that our curation effort lays the groundwork for a future overarching pathway ontology with cross-references to databases that could be leveraged and extended by the pathway community.

DATA AVAILABILITY

A freely available instance of DecoPath can be found at <https://decopath.scai.fraunhofer.de/>.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

We are very grateful to the curators of KEGG, Reactome, WikiPathways and PathBank for generating the raw content which was used in this work. Furthermore, we would like to thank Vasco Asturiano for developing *circle-packing*, the JavaScript library which is the basis of one of the visualizations of DecoPath.

Authors' contributions: D.D.F. conceived and designed the study. S.M. implemented the web application and analyzed the data with help from VSB and DDF. Y.G., S.M. and D.D.F. curated the pathway mappings. S.M. and D.D.F. wrote the paper. A.T.M., M.H.A., S.M. and D.D.F. acquired the funding.

All authors have read and approved the final manuscript.

FUNDING

This work was developed in the Fraunhofer Cluster of Excellence 'Cognitive Internet Technologies'.

Conflict of interest statement. D.D.F. received salary from Enveda Biosciences.

REFERENCES

- Dillies, M.A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J. *et al.* (2013) A comprehensive evaluation of normalization methods for illumina high-throughput RNA sequencing data analysis. *Brief Bioinform.*, **14**, 671–683.
- Reimand, J., Isserlin, R., Voisin, V., Kucera, M., Tannus-Lopes, C., Rostamianfar, A., Wadi, L., Meyer, M., Wong, J., Xu, C. *et al.* (2019) Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, cytoscape and enrichmentmap. *Nat. Protoc.*, **14**, 482–517.
- Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. *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.
- Nguyen, T.M., Shafi, A., Nguyen, T. and Draghici, S. (2019) Identifying significantly impacted pathways: a comprehensive review and assessment. *Genome Biol.*, **20**, 1–15.
- Bader, G.D., Cary, M.P. and Sander, C. (2006) Pathguide: a pathway resource list. *Nucleic Acids Res.*, **34**, D504–D506.
- Mubeen, S., Hoyt, C.T., Gemünd, A., Hofmann-Apitius, M., Fröhlich, H. and Domingo-Fernández, D. (2019) The impact of pathway database choice on statistical enrichment analysis and predictive modeling. *Front. Genet.*, **10**, 1203.
- Bateman, A.R., El-Hachem, N., Beck, A.H., Aerts, H.J. and Haibe-Kains, B. (2014) Importance of collection in gene set enrichment analysis of drug response in cancer cell lines. *Sci. Rep.*, **4**, 4092.
- Geistlinger, L., Csaba, G., Santarelli, M., Ramos, M., Schiffer, L., Turaga, N., Law, C., Davis, S., Carey, V., Morgan, M. *et al.* (2020) Toward a gold standard for benchmarking gene set enrichment analysis. *Brief Bioinform.*, **22**, 545–556.
- Zyla, J., Marczyk, M., Domaszewska, T., Kaufmann, S.H., Polanska, J. and Weiner, J. 3rd (2019) Gene set enrichment for reproducible science: comparison of CERNO and eight other algorithms. *Bioinformatics*, **35**, 5146–5154.
- Mathur, R., Rotroff, D., Ma, J., Shojaie, A. and Motsinger-Reif, A. (2018) Gene set analysis methods: a systematic comparison. *BioData Min.*, **11**, 1–19.
- Griss, J., Viteri, G., Sidiropoulos, K., Nguyen, V., Fabregat, A. and Hermjakob, H. (2020) ReactomeGSA-Efficient multi-omics comparative pathway analysis. *Mol. Cell Proteomics*, **19**, 2115–2124.
- Paczkowska, M., Barenboim, J., Sintupisut, N., Fox, N.S., Zhu, H., Abd-Rabbo, D., Mee, M.W., Boutros, P.C. and Reimand, J. (2020) Integrative pathway enrichment analysis of multivariate omics data. *Nat. Commun.*, **11**, 1–16.
- Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A.H., Tanaseichuk, O., Benner, C. and Chanda, S.K. (2019) Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat. Commun.*, **10**, 1–10.
- Canzler, S. and Hackermüller, J. (2020) multiGSEA: a GSEA-based pathway enrichment analysis for multi-omics data. *BMC Bioinform.*, **21**, 1–13.
- Stobbe, M.D., Houten, S.M., Jansen, G.A., van Kampen, A.H. and Moerland, P.D. (2011) Critical assessment of human metabolic pathway databases: a stepping stone for future integration. *BMC Syst. Biol.*, **5**, 165.
- Belinky, F., Nativ, N., Stelzer, G., Zimmerman, S., Iny Stein, T., Safran, M. and Lancet, D. (2015) PathCards: multi-source consolidation of human biological pathways. *Database*, bav006.
- Domingo-Fernández, D., Hoyt, C.T., Bobis-Álvarez, C., Marín-Llaó, J. and Hofmann-Apitius, M. (2018) ComPath: An ecosystem for exploring, analyzing, and curating mappings across pathway databases. *npj Syst. Biol. Appl.*, **4**, 43.
- Mora, A. and Donaldson, I.M. (2012) Effects of protein interaction data integration, representation and reliability on the use of network properties for drug target prediction. *BMC Bioinform.*, **13**, 1–17.
- Simillion, C., Liechti, R., Lischer, H.E., Ioannidis, V. and Bruggmann, R. (2017) Avoiding the pitfalls of gene set enrichment analysis with setrank. *BMC Bioinform.*, **18**, 151.
- Petri, V., Jayaraman, P., Tutaj, M., Hayman, G.T., Smith, J.R., De Pons, J., Laulederkind, S.J., Lowry, T.F., Nigam, R., Wang, S.J. *et al.* (2014) The pathway ontology—updates and applications. *J. Biomed. Semant.*, **5**, 7.
- Wishart, D.S., Li, C., Marcu, A., Badran, H., Pon, A., Budinski, Z., Patron, J., Lipton, D., Cao, X., Oler, E. *et al.* (2020) PathBank: a comprehensive pathway database for model organisms. *Nucleic Acids Res.*, **48**, D470–D478.
- Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M. and Tanabe, M. (2021) KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res.*, **49**, D545–D551.
- Fabregat, A., Korninger, F., Viteri, G., Sidiropoulos, K., Marín-García, P., Ping, P., Wu, G., Stein, L., D'Eustachio, P. and Hermjakob, H. (2018) Reactome graph database: Efficient access to complex pathway data. *PLoS Comput. Biol.*, **14**, e1005968.
- Martens, M., Ammar, A., Riutta, A., Waagmeester, A., Slenter, D.N., Hanspers, K., Miller, R.A., Digles, D., Lopes, E.N., Ehrhart, F. *et al.* (2021) WikiPathways: connecting communities. *Nucleic Acids Res.*, **49**, D613–D621.
- García-Campos, M.A., Espinal-Enríquez, J. and Hernández-Lemus, E. (2015) Pathway analysis: state of the art. *Front. Phys.*, **6**, 383.
- Khatri, P., Sirota, M. and Butte, A.J. (2012) Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput Biol.*, **8**, e1002375.

27. Xie,C., Jauhari,S. and Mora,A. (2021) Popularity and performance of bioinformatics software: the case of gene set analysis. *BMC Bioinform.*, **22**, 1–16.
28. Fisher,R.A. (1992) Statistical methods for research workers. *Breakthroughs in Statistics*. Springer, New York, NY, pp. 66–70.
29. Benjamini,Y. and Yekutieli,D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.*, **29**, 1165–1188.
30. The Cancer Genome Atlas Research Network, Weinstein,J.N., Collisson,E.A., Mills,G.B., Shaw,K.R.M., Ozenberger,B.A., Ellrott,K., Shmulevich,I., Sander,C. and Stuart,J.M. (2013) The cancer genome atlas pan-cancer analysis project. *Nat. Genet.*, **45**, 1113.
31. Colaprico,A., Silva,T.C., Olsen,C., Garofano,L., Cava,C., Garolini,D. and Noushmehr,H. (2016) TCGAAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.*, **44**, e71–e71.
32. Glaab,E., Baudot,A., Krasnogor,N., Schneider,R. and Valencia,A. (2012) EnrichNet: network-based gene set enrichment analysis. *Bioinformatics*, **28**, i451–i457.
33. Tarca,A.L., Draghici,S., Khatri,P., Hassan,S.S., Mittal,P., Kim,J.S., Kim,C.J., Kusanovic,J.P. and Romero,R., (2008) A novel signaling pathway impact analysis. *Bioinformatics*, **25**, 75–82.
34. Karp,P.D., Midford,P.E., Caspi,R. and Khodursky,A. (2021) Pathway size matters: the influence of pathway granularity on over-representation (enrichment analysis) statistics. *BMC Genomics*, **22**, 1–11.
35. Ai,C. and Kong,L. (2018) CGPS: a machine learning-based approach integrating multiple gene set analysis tools for better prioritization of biologically relevant pathways. *J. Genet. Genomics*, **45**, 489–504.
36. Alhamdoosh,M., Ng,M., Wilson,N.J., Sheridan,J.M., Huynh,H., Wilson,M.J. and Ritchie,M.E. (2017) Combining multiple tools outperforms individual methods in gene set enrichment analyses. *Bioinformatics*, **33**, 414–424.

A.4 Towards a global investigation of transcriptomic signatures through co-expression networks and pathway knowledge for the identification of disease mechanisms

Reprinted with permission from “Figueiredo R.Q., Raschka T., Kodamullil AT., Hofmann-Apitius M., Mubeen S.[†], and Domingo-Fernández D.[†] (2021). Towards a global investigation of transcriptomic signatures through co-expression networks and pathway knowledge for the identification of disease mechanisms. *Nucleic acid research*, 49(14): 7939–7953”.

Copyright © Figueiredo R.Q., *et al.*, 2021.

Towards a global investigation of transcriptomic signatures through co-expression networks and pathway knowledge for the identification of disease mechanisms

Rebeca Queiroz Figueiredo^{1,2,†}, Tamara Raschka^{1,2,3,†}, Alpha Tom Kodamullil^{1,4}, Martin Hofmann-Apitius^{1,2}, Sarah Mubeen^{1,2,3,†} and Daniel Domingo-Fernández^{1,3,5,*†}

¹Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing, Sankt Augustin 53757, Germany, ²Bonn-Aachen International Center for IT, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn 53115, Germany, ³Fraunhofer Center for Machine Learning, Germany, ⁴Causality Biomodels, Kinfra Hi-Tech Park, Kalamassery, Cochin, Kerala, India and ⁵Enveda Biosciences, Boulder, CO 80301, USA

Received March 18, 2021; Revised May 17, 2021; Editorial Decision June 08, 2021; Accepted June 11, 2021

ABSTRACT

We attempt to address a key question in the joint analysis of transcriptomic data: can we correlate the patterns we observe in transcriptomic datasets to known interactions and pathway knowledge to broaden our understanding of disease pathophysiology? We present a systematic approach that sheds light on the patterns observed in hundreds of transcriptomic datasets from over sixty indications by using pathways and molecular interactions as a template. Our analysis employs transcriptomic datasets to construct dozens of disease specific co-expression networks, alongside a human protein-protein interactome network. Leveraging the interoperability between these two network templates, we explore patterns both common and particular to these diseases on three different levels. Firstly, at the node-level, we identify most and least common proteins across diseases and evaluate their consistency against the interactome as a proxy for their prevalence in the scientific literature. Secondly, we overlay both network templates to analyze common correlations and interactions across diseases at the edge-level. Thirdly, we explore the similarity between patterns observed at the disease-level and pathway knowledge to identify signatures associated with specific diseases and indication areas. Finally, we present a case scenario in schizophrenia, where we show how our approach can be used to investigate disease pathophysiology.

INTRODUCTION

Despite the exponential growth of biomedical data in the last decades, we are still far from understanding the function of every gene in a living organism. Nevertheless, major technological advancements now enable us to assign specific biological functions to thousands of protein-coding genes in the human genome (1). In turn, complex interactions between groups of genes, proteins and other biomolecules give rise to the normal functioning of the cell. By acquiring knowledge of these interactions, we can decipher the molecular mechanisms which cause system-wide failures that can lead to disease (2). A common modeling approach to represent these vast sets of interactions is in reconstructing mechanisms in the form of networks as intuitive representations of biology, where nodes denote biological entities and edges their interactions (3,4).

Numerous standardized formats have been widely adopted to model biological networks that represent pathway knowledge dispersed throughout the scientific literature (5). Pathway models in a variety of formats can be found housed in databases such as KEGG (6) and Reactome (7), each with a varied focus and scope. These databases can be specifically leveraged for hypothesis generation, the analysis of biomedical data such as with pathway enrichment (8), or predictive modeling (9). Using the networks of known molecular interactions, one can also discern novel genes involved in particular disease states as functions of network proximity (10). A general trend noted by Huang and colleagues was the observation that larger networks tended to outperform smaller ones, an effect also observed when comparing the performance of integrated pathway databases to individual ones in enrichment and predictive modeling tasks (11).

*To whom correspondence should be addressed. Tel: +49 2241 14 4036; Email: daniel.domingo.fernandez@scai.fraunhofer.de

†The authors wish it to be known that, in their opinion, the first two and last two authors should be regarded as Joint First/Last Authors.

Although knowledge-driven approaches that leverage literature-based evidence can be used to gain a mechanistic understanding of disease pathophysiology, these approaches tend to be augmented when applied in combination with data-driven ones. In the latter case, transcriptomic profiling offers researchers a systematic and affordable method to analyze the expression and activity of genes and proteins on a large-scale under distinct physiological conditions. Through gene expression profiling, patterns of genes expressed at the transcript level that are relevant to a particular condition can be determined, whilst considering sets of genes involved in a specific biological process tend to exhibit similar patterns of expression or activity (12). To model these patterns, techniques such as gene co-expression networks have been developed in which genes with correlated expression activity are connected. Several methodologies can be used to generate co-expression networks, such as Weighted Gene Co-expression Network Analysis (WGCNA) (13), SWItchMiner (SWIM) (14) and ARACNE (15). Co-expression networks tend to be represented as undirected weighted graphs, where graph nodes correspond to genes, and edges between nodes correspond to co-expression relationships (16). The applications of these networks are diverse, ranging from identifying functional and disease-specific modules to hub genes (12). For instance, Chou *et al.* (17) and Xiang *et al.* (18) combined independent datasets related to endometrial cancer and Alzheimer's disease, respectively, in order to generate co-expression networks that captured gene expression patterns across multiple disease-specific datasets. Using these co-expression networks, they were able to identify relevant genes in the context of these two indications.

Though it is standard practice to perform enrichment analysis using pathway and gene set databases (e.g. KEGG and Gene Ontology (19)) on gene lists from co-expression networks such as those from a particular disease module (20,21) for mechanistic insights, this approach ignores the topology of the network as it exclusively relies upon sets of genes rather than the network structure. In a recent study, Paci *et al.* (22) overcame this challenge by showing how distinct, topological properties of disease networks can emerge through the identification and mapping of disease-specific genes of several disease co-expression networks to a human interactome network of protein-protein interactions. The SWIM method used by the authors has independently been applied to elucidate the molecular mechanisms that underlie several complex diseases mediated by the identification of key genes (23–26).

The potential insights that can be gained from the previously mentioned analyses together with the abundance of publicly available transcriptomic datasets (27,28) have prompted the creation of databases that store collections of co-expression networks, such as COXPRESdb for numerous species (29). By harmonizing and storing thousands of transcriptomic datasets in the form of co-expression networks, these resources capture a variety of ‘snapshots’ representing gene expression patterns in a diverse set of contexts. While transcriptomics datasets have been used to identify regulatory patterns across a variety of different contexts such as specific species or tissues (30), the aim of most transcriptomics data analyses is to reveal biological pro-

cesses that differentiate a disease patient from a healthy control. The large amount of datasets available contain an abundant number of samples, allowing for comprehensive large-scale analyses on a variety of indications. Furthermore, by bringing together transcriptomic data with known interactions in pathway resources, we can connect the transcriptome with the proteome by overlaying the patterns in co-expression networks with the scaffold of biological knowledge embedded in pathway networks (31). In doing so, we can gain insights on specific or shared molecular mechanisms across multiple indications.

In this work, we jointly leverage the patterns of disease-specific datasets reflected in co-expression networks and pathway and interaction networks to uncover the mechanisms underlying disease pathophysiology. To do so, we systematically compared hundreds of transcriptomic datasets from over 60 diseases with a human protein-protein interactome network to unravel the proteins, subgraphs, and pathways that are specific to certain diseases or shared across multiple. Finally, in a case scenario, we demonstrate how bringing together a disease-specific co-expression network with pathway knowledge allows us to better understand the role of a specific pathway within a disease context.

MATERIAL AND METHODS

In the first subsection, we outline the process of generating disease-specific co-expression networks from transcriptomic data (Figure 1A–C). Then, we describe the construction of a human protein-protein interactome network (Figure 1E and F). Finally, we outline the various analyses conducted (Figure 1D).

Generating co-expression networks from transcriptomic data

Identifying disease-specific datasets in ArrayExpress. We queried datasets from ArrayExpress (AE) (27) belonging to the most widely used platform: the Affymetrix Human Genome U133 Plus 2.0 Array (accession on AE: A-AFFY-44). By using the same platform for each of the datasets, we ensured that the datasets were relatively comparable. ArrayExpress was preferred over other databases such as Gene Expression Omnibus (GEO) (28) as datasets often comprise of normalized and mapped terms in their metadata that describe their characteristics (e.g. experimental details, organism information, etc.). Furthermore, it provides a user-friendly API through which all the necessary information was queried. As of 20 July 2020, 4485 datasets generated from platform A-AFFY-44 have been stored in ArrayExpress, resulting in roughly below 200 000 samples. Figure 2 summarizes the filtering steps that we conducted to identify disease-specific datasets which are also described below.

As the purpose of this work was to analyze disease-specific datasets, only patient samples and their controls were eligible for the analysis. Thus, a filtering step was introduced to focus exclusively on patient-level data (Figure 2, filter A). To filter out irrelevant datasets, we leveraged keywords present in the metadata such as ‘dose’, ‘compound’ or ‘strain’ (Figure 2, filter B). Furthermore, information about the disease state of each sample is needed for building disease-specific networks. Therefore, the metadata columns

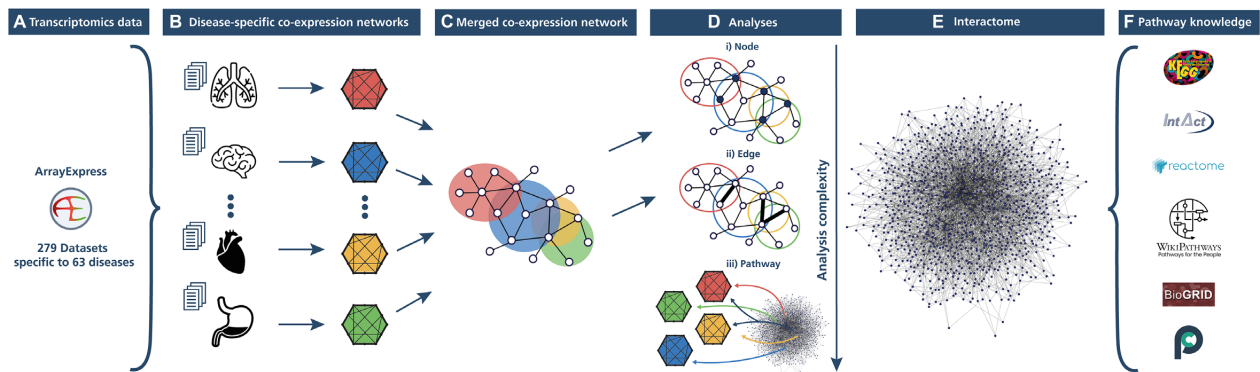


Figure 1. Schematic illustration of the methodology. 279 transcriptomic datasets were acquired from ArrayExpress (A) and grouped into 63 distinct diseases to generate disease-specific co-expression networks (B). A comprehensive protein-protein interactome network was built (E) from an ensemble of six pathway and interaction databases (F). A series of analyses were then conducted on the disease-specific co-expression networks (D), specifically: a node-level analysis (D.i), an edge-level analysis (D.ii), and a pathway-based analysis (D.iii) leveraging pathway knowledge and the interactome network.

were searched for disease keywords like ‘disease’, ‘histology’ or ‘status’ (Figure 2, filter C). This resulted in 651 datasets, of which one non-human dataset was removed, totaling in 650 datasets with 51 550 samples (Figure 2, filter D).

Once datasets were filtered to identify those that contained disease-specific information, we then harmonized the disease terms present in the title and metadata of the datasets with the help of the Human Disease Ontology (DOID) (32). Next, the disease terms from patient samples were mapped to DOID entities using ZOOMA (<https://www.ebi.ac.uk/spot/zooma>), enabling us, in some cases, to automatically find DOID matches. However, the majority of the terms did not contain a perfect match to a DOID entity so ZOOMA proposed the closest match. Based on this set of proposed DOID entities, we manually evaluated whether the term had been correctly assigned or if a DOID entity that could more accurately represent the disease was available. Through this process, we also identified false positive terms which had not been successfully filtered in the previous steps. In these cases, the metadata did not contain sufficient information, though this information was present in the dataset title. Thus, using the title information, we removed such false positive terms following manual inspection.

To maximize the coverage, we conducted a final processing step where we intended to group similar diseases together under a common label. For that, we leverage the ontology network structure and visualize it as a hierarchical tree with a focus on selected branches (i.e. ‘immune system disease’, ‘nervous system disease’, and ‘cancer’). Next, we manually identify close neighbors for terms that have few samples in order to merge them into a more general term that still accurately describes the original term. The veracity of the likelihood of the disease terms in the selected clusters to be used as a single gene expression set were verified by a clinician before re-mapping. (Supplementary Text 1).

After this final grouping step, we also filtered datasets to fulfill the following criteria: i) ensure every disease has a minimum of 50 samples to increase the stability of the co-expression network, ii) ensure a minimum of two datasets per disease, and iii) exclude samples with the ‘cancer’ la-

bel as this term was too broad (Figure 2, filter F). Thus, we have 38 621 samples from 469 datasets as 63 distinct diseases and one control group (Supplementary Table S1). To facilitate the grouping of control samples, we first harmonized all samples coming from datasets used to generate the disease networks that correspond to controls by giving them a common label (i.e. ‘normal’) (Figure 2, filter G). Applying the previously described filtering steps resulted in 35 025 samples from 323 datasets that were selected. Finally, not all datasets comprised the raw data required to generate the co-expression networks which are solely based on 279 datasets (20 748 samples) (Figure 2, filter H). The final list of datasets with their respective disease labels can be seen in Supplementary Table S1 and can be visualized according to their DOID hierarchy in Supplementary Figure S1.

Scripts to retrieve and process the datasets from ArrayExpress are available at <https://github.com/CoXPath/CoXPath/blob/main/R>. We have also provided comprehensive documentation to modify the filtering steps and add extensions to the scripts.

Generating co-expression networks. For each disease, expression data could then be used to construct co-expression networks to represent relationships between genes in different diseases. Therefore, the raw .CEL-files of the expression datasets were downloaded, pre-processed, and merged. Here, each individual dataset was first pre-processed with the RMA function of the oligo package in R, which performs background subtraction and quantile normalization. After merging the samples from different datasets irrespective of the sample tissue (as this information was not available for a large amount of samples), a batch correction via ComBat (33) was applied to the data to remove the effect corresponding to individual datasets. Finally, the probes were mapped to genes. If multiple probes mapped to the same gene, the most variable probe was kept. In the special case of the normal network, we would like to note that only control samples that were present in the disease datasets were used (Figure 2, filter G).

The actual co-expression datasets were then constructed with the WGCNA package in R (13). WGCNA has been shown to be one of the most accurate methods, even in the

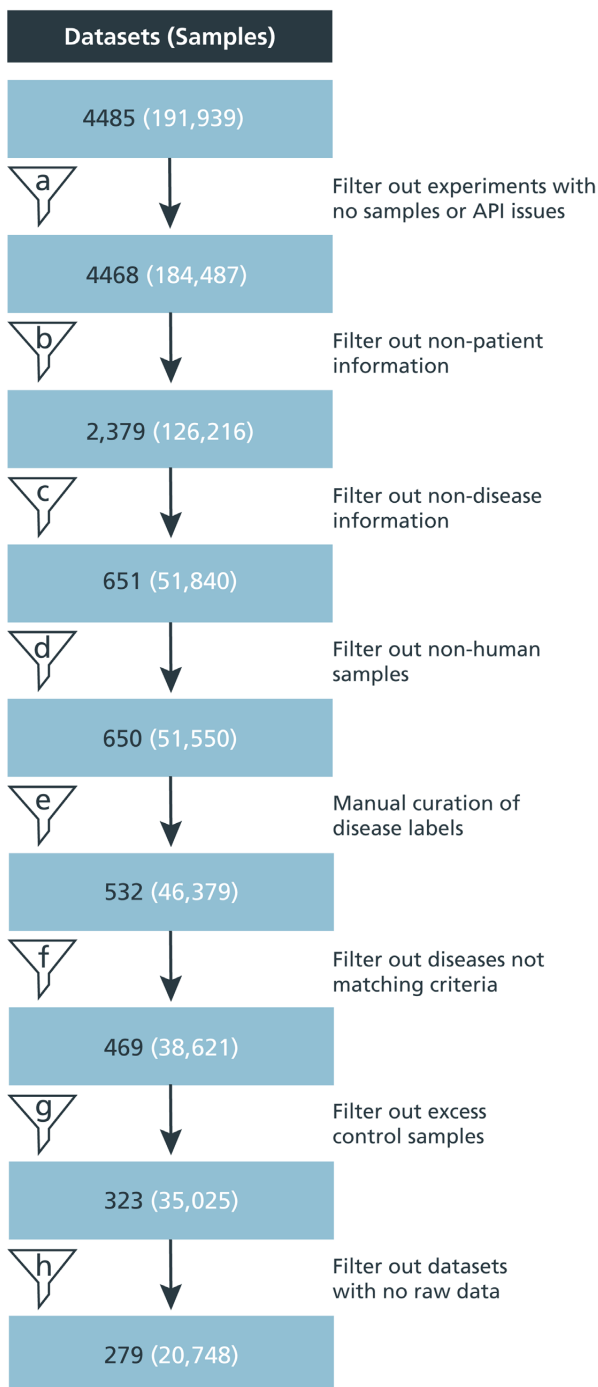


Figure 2. Extracting disease-specific datasets from ArrayExpress. transcriptomic data from nearly 4500 datasets was derived from ArrayExpress. Several filtering steps (A–H) were applied to only retain disease-specific datasets for patient samples and controls that fulfilled the criteria outlined in the section: Identifying disease-specific datasets in ArrayExpress.

case of small sample sizes, as opposed to other methods such as ARACNE (34). But, in contrast to most common approaches that construct and analyze modules of the network based on hierarchical clustering, here we relied only on the topological overlap matrix (TOM). In order to fa-

cilitate the comparability of the networks, for each disease, we defined its co-expression network as the top 1% highest similarity in the TOM as it is considered a conservative cut-off in benchmark studies (35) and enables us to maintain the same number of edges in each network while the number of nodes can vary. Nodes having a higher topological overlap were previously found to be more likely to belong to the same functional class than nodes having lower topological overlap (36). Given the platform used in this study (i.e. Affymetrix Human Genome U133 Plus 2.0 Array), 1% corresponds to 2 036 667 edges for each co-expression network. This cut-off for connections with the highest topological overlap was used because without a stringent cut-off, we would yield fully connected networks of over 200 million edges. However, since most of the genes do not have such a high topological overlap, the majority of these edges would not be relevant for our analysis as they would have a weight close to zero (see Supplementary Text 2 for more details). Finally, we would like to mention that we refer to these edges interchangeably as correlations through this paper. Although this is not precise, edges representing a high topological overlap are also highly correlated as the TOM value is based on the signed correlation but also takes the connectedness of nodes into account.

Building a human protein–protein interactome network

To systematically compare disease-specific co-expression networks against pathway knowledge, we built an integrative network comprising information from a compendium of well-established databases. This interactome was comprised of tens of thousands of human protein–protein interactions from six databases including KEGG (6), Reactome (7), WikiPathways (37), BioGrid (38), IntAct (39) and PathwayCommons (40). We would like to note that the first three of the six databases were harmonized through PathMe (41). Additionally, for each of the six databases, only proteins that belonged to pathways from MPath (11), an integrative resource that combines multiple databases and merges gene sets of equivalent pathways, were included in the interactome, thus ensuring that each protein in the network was minimally assigned to a single pathway. The use of MPath to annotate proteins to pathways facilitated both the generation of a larger network and the avoidance of redundant pathways.

The resulting human interactome has a total of 8601 nodes and 199 535 edges. Not surprisingly, the vast majority of the nodes in the interactome are protein-coding genes, as these genes are transcribed into functional proteins with essential roles in the biological processes represented in pathway databases (Figure 3A). Among the edges of the interactome, association relations are the most prevalent (~73%), while causal relations including, increase, decrease, regulate, and has_component relations constitute the remaining relation types (Figure 3B). Apart from the interactome network we generated, we also obtained protein–protein interactions (PPIs) from HIPPIE (42) and STRING (43) to compare the results yielded by our network containing pathway knowledge with other comprehensive PPI resources. Unlike the protein–protein interactome, the STRING and HIPPIE networks were not constrained to only contain pro-

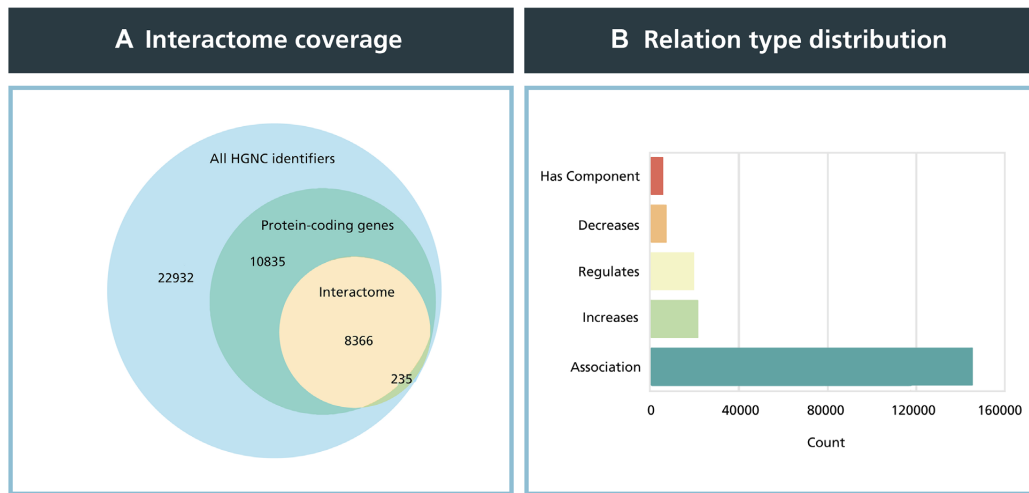


Figure 3. Node and edge type statistics of the human protein-protein interactome network. **(A)** Venn diagram indicating the coverage of proteins in the interactome network with respect to all existing HGNC identifiers as well as protein-coding genes. The interactome contains ~8600 unique HGNC identifiers, or 20% of the roughly 42 300 approved HGNC identifiers. In total, 97% of the HGNC identifiers of the interactome are protein-coding genes. **(B)** Distribution of relation types in the interactome network. The largest proportion of relation types were associations, comprising nearly 73% of all ~200 000 edges, while causal relations, specifically decrease, regulate, and increase, made up ~25% of all relation types with roughly 50 000 edges.

teins which could be annotated to pathways. Thus, both networks contained a significantly greater number of proteins and PPIs (detailed descriptions of these networks can be found in Supplementary Table S2).

Analyses

Software and data used in network analysis and visualization. Network analyses were conducted using the methods and algorithms implemented in NetworkX (v2.5) (44). KEGG pathways (6) were downloaded on 3 August 2020 using ComPath (41). Network visualizations were done using WebGL, D3.js, and Three.js and the python-igraph package. The processed data and analyses are available at <https://github.com/CoXPath/CoXPath>.

Meta-analysis of gene expression data. Differential expression analysis was performed using the Limma R package (45) on the merged disease datasets previously described which contained information on both patient samples and controls. This step yielded differentially expressed genes (DEGs) for 46 diseases in total from the original 63. For all other diseases, no matching control data was available (Supplementary Table S3). For example, both datasets (E-GEOD-13141 and E-GEOD-16237) used to build the neuroblastoma dataset only consisted of neuroblastoma samples. DEGs for each disease were then filtered to include only those with an adjusted P -value < 0.05 . DEGs across the 46 diseases were combined into a consensus by splitting the up- and down-regulated genes for each disease and taking the average adjusted P -values and \log_2 fold changes for all up- and down-regulated subsets separately.

Quantifying the similarity between disease-specific co-expression networks and biological pathways. To investigate the consensus between the patterns present in each co-expression network and pathway knowledge,

we superimposed each disease-specific co-expression network against pathways from KEGG and the interactome network using two different methods. Method 1 investigates every pairwise combination of nodes from the set of proteins P for a given pathway from KEGG (C_P) to find the proportion of edges that exist in the disease co-expression network $D = (P', E_D)$ between those node pairs, namely edge overlap (edge overlap = $|\{\forall e_{u,v} \text{ s.t. } u, v \in C_P; u, v \in P' \text{ and } e_{u,v} \in E_D\}|$) (Equation 1). P' is the set of proteins in the co-expression network and E is the set of edges connecting the proteins.

$$\text{pathway-based similarity } (P, D) = \frac{\text{edge overlap}}{|C_P|} \quad (1)$$

Equation 1. Similarity between a pathway and disease co-expression network using method 1.

Similarly, applying a more stringent criterion to take into account the protein-protein interactome network, using a set of proteins P for a given pathway from KEGG, method 2 takes the interactome network $I = (U, E_i)$ and generates a subgraph $S = (V, E_S)$ containing only those nodes in P with edges in E_i (with $V = \{u : e_{u,v} \in E_S \text{ and } u, v \in P \cap U\}$ and $E_S = \{e_{u,v} : u, v \in P; u, v \in U; e_{u,v} \in E_i\}$). Next, the proportion of edges on the interactome subgraph S that are also found in each disease co-expression network $D = (P', E_D)$ are calculated (Equation 2).

$$\text{interactome-based similarity } (P, D) = \frac{|E_S \cap E_D|}{|E_S|} \quad (2)$$

Equation 2. Similarity between a pathway and disease co-expression network using method 2.

We would like to mention that we exclusively used pathway definitions (i.e. gene sets) from KEGG which contain a relatively fewer number of pathways in order to facilitate the interpretation of the analysis (e.g. Reactome contains

over 2000 pathways while KEGG has over 300). Nonetheless, in method 2, we overlay the KEGG gene sets onto the interactome network, ensuring that the analysis is not only restricted to biological interactions in KEGG.

Pathway enrichment analysis. Overrepresentation analysis (ORA) was conducted employing a one-sided Fisher's exact test (46) for each of the pathways in KEGG (downloaded on 12 December 2020). A pathway is considered to be significantly enriched if its adjusted *P*-value is smaller than 0.05 after applying multiple hypothesis testing correction using the Benjamini–Yekutieli method (47).

RESULTS

In the first subsection, we outline the diseases that fulfilled the criteria to generate the corresponding co-expression networks and investigate the characteristics of these networks. Then, we analyze the disease-specific co-expression networks at the node- and edge- levels, respectively, while later comparing the co-expression networks against pathway knowledge. Finally, in a case scenario we demonstrate how a pathway-level analysis in a disease context can be leveraged to better understand the role of a specific pathway in a disease context.

Overview of disease-specific co-expression networks

From over 330 datasets that were categorized into distinct diseases, we systematically constructed 64 co-expression networks, 63 of which correspond to disease-specific co-expression networks, and the remaining corresponding to a control group co-expression network. Figure 4A summarizes the network size of each disease-specific co-expression network clustered by major disease indication for a total of ten disease categories and one unspecific group. Body system clusters (e.g. gastrointestinal system disease, immune system disease) were given priority for the classification of all cancers before considering the 'other cancer' group. How each disease relates to its disease category cluster can be visualized on the DOID hierarchy in Supplementary Figure S1. The sarcoma co-expression network had the least number of nodes of all the networks (i.e. 5450), while the ductal carcinoma in situ co-expression network had the highest number of nodes (i.e. 20 163). Generally, the networks within each disease category cluster tended to vary greatly in size. For example, the 'immune system disease' category includes networks ranging in size from 5754 to 18 449 nodes. Additionally, the number of co-expression networks within a disease cluster varied, with nearly half the disease groups containing between 6 and 15 networks (i.e. gastrointestinal system disease, immune system disease, nervous system disease, respiratory system disease, and other cancer), while all remaining clusters contained less than five.

We also investigated whether a correlation exists between the number of samples or datasets used to create a co-expression network and the size of the network. No dependency of network size based on the amount of samples/datasets used was observed (Supplementary Figure S2). The total number of datasets ranged from 1 to

27, while the total number of samples was between 9 and 2515. The vast majority of disease co-expression networks were generated from 1 to 10 datasets and contained between 9 to 461 samples. We found that the resulting network size for each disease varied within a wide range (i.e. between ~6000 and 20 000) and no discernible pattern was observed.

Investigating global trends of disease-specific co-expression networks at the node level

Exploring the most and least common proteins of the co-expression networks. Here, we explored the most and least common proteins across all 63 disease-specific co-expression networks generated with the goal of identifying both disease-specific and commonly occurring proteins. We first identified the most common proteins as those that occur in the highest number of disease co-expression networks (Supplementary Figure S3). We discovered that 96–99% of the top 1000 to top 100 most common disease proteins, in intervals of 100, are also found in the normal network, indicating that these proteins are widespread and therefore not disease-associated proteins. Additionally, we found that none of the proteins were present in all co-expression networks as we were only interested in considering the top 1% strongest correlations in each network (i.e. the selected cut-off; see Generating co-expression networks section). On the other hand, TXLNGY and NCR2 were the most common proteins, occurring in 60 out of the 63 disease co-expression networks. Nonetheless, we were able to identify 48 proteins present in at least 57 of the 63 diseases.

We next assessed the overlap between all proteins of the interactome and the disease co-expression networks (Supplementary Figure S4). From this overlap, we investigated whether proteins in the disease co-expression networks could consistently be identified in our interactome network to infer how well these proteins have been studied and reported in the literature. We refer to proteins that could consistently be found across the majority of disease co-expression networks and were also present in the interactome as the most common proteins of the disease networks and the most highly connected proteins of the interactome. Surprisingly, we found that only 30–33.4% of the most common proteins (with cut-offs between 50 and 54 out of 63) of the disease co-expression networks were present in the interactome. Similarly, for an approximately proportional range of these most common disease proteins against the most connected proteins of the interactome (i.e. top 100–400 proteins), little to no overlap was observed (Supplementary Figure S5). We also found that the average number of relations for the proteins in the interactome that overlapped with the approximately top 400 most common proteins in the disease networks (~33 relations) was lower than the average number of relations overall in the interactome (~46 relations). This analysis was also conducted on networks built using the STRING and HIPPIE PPI resources; relative to the interactome, we found a much larger overlap between proteins in these networks and proteins of the disease co-expression networks (Supplementary Figures S6 and S7). Within these overlaps, we observed similarly small overlaps of the most com-

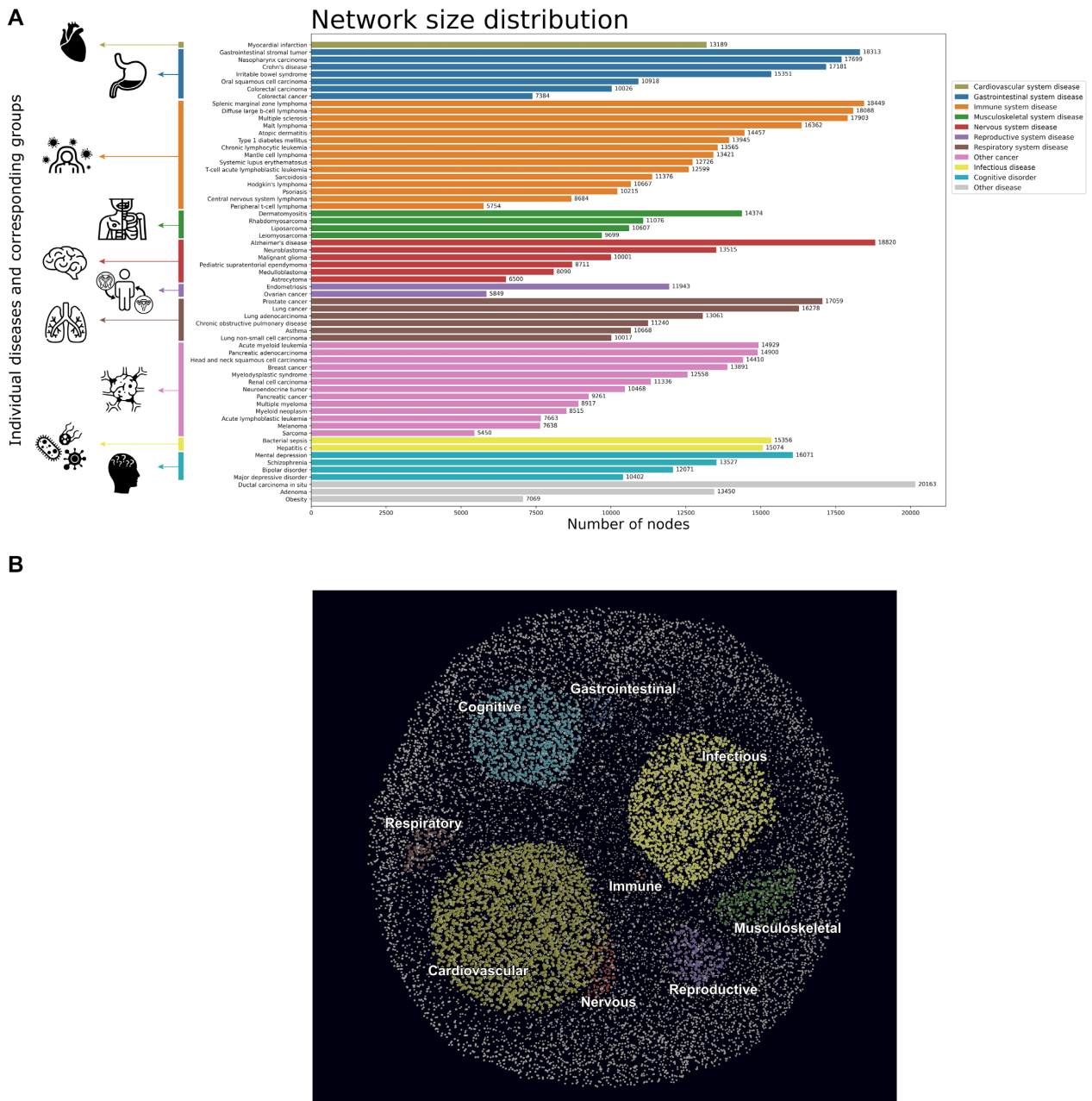


Figure 4. (A) Overview of the size of each of the co-expression networks clustered by major disease groups. (B) Merged co-expression network clustering proteins by their association to different disease groups. In A each of the 63 diseases was grouped into one of ten categories (or a remaining leftover group). Here, we see the varied sizes of co-expression networks within their corresponding disease clusters. In B association was determined by selecting the set of nodes which were present in all of the diseases of a given disease cluster (excluding ‘other’ and ‘other cancer’), and eliminating those nodes which were also present in all diseases of other clusters. This resulted in unique sets of nodes which were guaranteed to be found in all diseases of the given cluster, but not in all of another cluster. As expected, we observed an inverse correlation between the number of diseases in a cluster and the size of the associated node subset. High quality versions are available at <https://github.com/CoXPath/CoXPath/tree/main/results/figures>.

mon disease proteins and the most connected proteins of these two networks (Supplementary Figures S8 and S9). We then evaluated the overlap between all proteins from KEGG pathways and the disease co-expression networks (Supplementary Figure S10) and sought to verify whether the most common proteins in the disease co-expression networks could also be found in pathway databases. In doing

so, we identified only a small proportion (i.e. 29–31%) of these proteins in KEGG (Supplementary Table S4). When comparisons were made against KEGG pathway annotations, we observed that these few most common proteins had, on average, a slightly lower number of pathway annotations (~14.8) than the average number of annotations for all proteins in the pathway database (~16). Taken to-

gether, these findings indicate that though these proteins are the most common across all disease co-expression networks, they tend to be underrepresented in the scientific literature.

Among the proteins in common between the top 400 most highly connected proteins of the interactome and the most common proteins in the disease co-expression networks, 13 proteins, including three members of the cytochrome P450 family of enzymes (i.e. CYP1A2, CYP2C9 and CYP3A4), a major ribosomal protein (i.e. RPL18), as well as key regulatory proteins such as CDK1, PRKCG and PLCB2 were present in the overlap (Supplementary Table S4 and Supplementary Figure S6). Similarly, we define proteins that could be consistently found across the majority of disease co-expression networks and were also present in KEGG pathway annotations as the most common proteins of the disease networks and the most common KEGG proteins. We examined the overlap between the top 400 most common disease proteins with the highest number of KEGG pathway annotations, and the most common proteins of the disease co-expression networks (Supplementary Figure S11). Among the 22 proteins in common, we found seven members of the human leukocyte antigens (HLA) system of proteins (HLA-B, HLA-C, HLA-DMA, HLA-DMB, HLA-DQB1, HLA-DRA and HLA-G), as well as several proteins which were also in the overlap between the aforementioned most highly connected proteins of the interactome and most common proteins in the disease co-expression networks (i.e. CAMK2A, ELK1, GNAO1 and PLCB2) (Supplementary Table S4 and Supplementary Figure S11).

Finally, we investigated the least common proteins in the disease co-expression networks and their overlap with those in the interactome, additional PPI resources and pathway knowledge (Supplementary Figures S12–S15). Similar to the most common ones, we found that the majority of the least commonly occurring proteins in the co-expression networks were not present in the interactome nor in KEGG, suggesting that little is currently known of these proteins. Among the least commonly occurring proteins that overlapped with proteins from both KEGG and the interactome, we observed a significant number of proteins from the ZNF family (i.e. 42/54 (78%) from KEGG and 12/43 (28%) from the interactome overlap) (Supplementary Table S4). This family is one of the largest protein families and is known to regulate a wide range of biological processes, while some of its members have already been associated with several disorders (48). Thus, it may be interesting to investigate proteins that are specific to a particular disease, or a few distinct diseases, in detail. As an example, we observed that TWIST1, one of the least commonly occurring proteins and a well-known oncogene (49), was exclusively present in only 25 diseases and over 50% of them were cancers (Supplementary Table S4 and Supplementary Figure S15).

Meta-analysis on consistently differentially expressed genes across diseases. Differential gene expression analysis was performed in order to pinpoint genes which were consistently significantly differentially expressed between patient and control samples across 46 diseases. While here, we in-

dependently conducted the meta-analysis to identify patterns of dysregulation among DEGs that are specific to or shared across diseases, in the case scenario we demonstrate how DEGs can be overlaid with disease-specific co-expression networks and the interactome to elucidate mechanisms that DEGs may be involved in. The average of all genes in these diseases that were up-regulated as well as the average of all genes that were down-regulated were independently calculated. Figure 5 jointly reports the comparison of the negative \log_{10} adjusted P -values versus \log_2 fold changes of all independently averaged up- and down-regulated DEGs in the 46 diseases. We found that nearly all genes were, to some degree, up-regulated in one or more diseases and down-regulated in at least one other, while only CCDC43, JADE3, RPL22L1, SOCS1 and TOR3A were exclusively up- and CAVIN2 and ZSCAN18 down-regulated across all diseases they were present in. In all, nearly 20 000 unique genes were significantly differentially expressed (adjusted P -value < 0.01), with ~ 17 600 up-regulated DEGs and ~ 15 600 down-regulated ones.

We then applied a \log_2 fold change threshold of 1.75 to identify significantly (adjusted P -value < 0.01) differentially expressed genes with the most extreme average \log_2 fold change values. This threshold was selected as it yielded a reasonable number of DEGs to investigate (i.e. 60), whereas more commonly used thresholds, such as \log_2 fold change > 1.5, yielded over 200. Among the genes that were found to be significantly differentially expressed at the extremes, 34 were the most up-regulated and 26 were the most down-regulated (Supplementary Table S5).

These genes were then compared to the top 500 most and least common disease proteins. Of the genes that were the most up-regulated, CDK1 was also among the top 500 most common disease proteins, while CRNDE, DEPTOR, and RASD1 were among the 500 least common. Similarly, for genes that were the most down-regulated, only S100A8 was among the top 500 most common disease proteins while no genes overlapped with the 500 least common disease proteins. Additionally, we found that four of the most up-regulated genes belonged to the collagen group of protein (i.e. COL11A1, COL1A1, COL1A2 and COL3A1), while some protein families (i.e. S100 protein family, SLC, and SYNP) could be found both in the most up- and down-regulated genes.

Of the most significantly highly up- and down-regulated genes (i.e. adjusted P -value < 0.01; \log_2 fold change > 1.75), we examined their expression changes in each of the individual diseases they were involved in. Interestingly, we found a group of genes (i.e. AMPD1, BEX5, DEPTOR, IGF1, JCHAIN, MARC2, MTUS1 and NDFIP2) that were highly up-regulated in only two of nearly 20 diseases they were in (i.e. myeloid neoplasm and multiple myeloma, grouped in the other cancers cluster), and down-regulated in nearly all of the remaining. Thus, although these genes were down-regulated in the vast majority of diseases they were involved in, they still appeared among the most significantly highly up-regulated genes since they are significantly up-regulated in the two aforementioned cancers. This trend has been documented for DEPTOR, with low expression of the gene observed in most cancers, yet high overexpression seen in a group of multiple

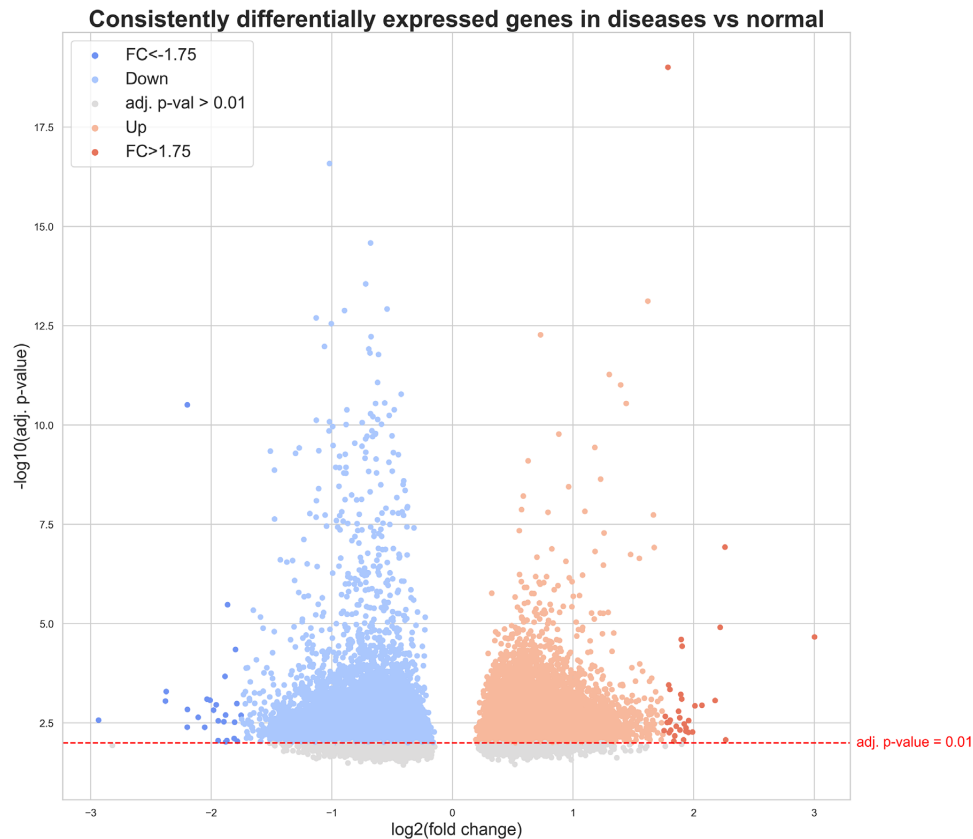


Figure 5. Consensus for consistently differentially expressed genes. Genes for 46 diseases were split into two subsets: those that were up-regulated and those that were down-regulated in that disease. The average consensus was taken for all up- and down-regulated subsets separately and is shown here. Nearly all genes could be found in both the up-regulated and down-regulated consensus as most genes are up-regulated in at least one disease as well as down-regulated in at least one other disease. In total, 19 666 unique genes were significantly up- or down-regulated (i.e. adjusted P -value < 0.01 ; above red line). Of the significantly differentially expressed genes, 17 643 genes were up-regulated and 15 634 genes were down-regulated. The most significantly differentially expressed genes were defined as additionally having a \log_2 fold change > 1.75 , resulting in 26 most down-regulated genes (dark blue) and 34 most up-regulated genes (dark orange).

myelomas (50). Similarly, among the genes highly down-regulated, we identified a subset of genes (i.e. *ASH1L-AS1*, *CXCL8*, *DUSP4*, *EPC1*, *PANK2*, *PCIF1*, *PHLDA1* and *PMAIP1*) that were only highly down-regulated in peripheral T-cell lymphoma, whilst being up-regulated in nearly all of the remaining diseases they were in (i.e. 17 diseases on average). This pattern has been identified with the over-expression of *DUSP4*, a tumor suppressor, in certain cancer types (51), whereas the loss of its expression caused by epigenetic dysregulation has been observed in at least one type of lymphoma (52) Finally, the meta-analysis revealed that one gene with significantly highly down-regulated, *SLC8A1*, was only significantly down-regulated in a group of nervous system diseases (i.e. medulloblastoma, pediatric supratentorial ependymoma, malignant glioma, astrocytoma, and to a lesser degree, Alzheimer's disease), not altogether surprising as the *SLC8* gene family of sodium-calcium exchangers, which includes *SLC8A1*, have been shown to play important regulatory roles in the control of central nervous system functions (53). In contrast, *SLC8A1* was only identified as significantly up-regulated in multiple myeloma.

Investigating global trends of disease-specific co-expression networks at the edge level

In this subsection, we explored the most commonly occurring edges among the co-expression networks from all diseases and compared them against the normal co-expression network and the interactome. We hypothesized that the edges most common across disease networks involve the dysregulation of key proteins such as transcription factors, common edges present across several diseases as well as in the normal co-expression network correspond to non-specific interactions, and common edges which could also be identified in the interactome represent known molecular interactions.

We first assessed whether there were any edges specific to particular disease networks, identifying 57 774 118 unique edges in total (i.e. 45% of all edges). This was to be expected, as we exclusively focused on the 1% strongest correlations from the initial hundreds of millions of possible edges, which led to most of the edges in our resulting co-expression networks to be specific to a single disease. Although this unique, disease-specific set of edges are worth exploring, due to the considerably large number of edges

in the co-expression networks, we restricted our analysis to the most common edges in the co-expression networks. We found that 21 edges were in more than 70% of the diseases (44/63) and 202 in more than 50% of the diseases (32/63). Interestingly, of those 21 edges that were in 70% of the diseases, we observed that 6 of the 13 proteins which are encoded by genes in the Y chromosome appeared in 5 edges each (i.e. RPS4Y1, USP9Y, DDX3Y, KDM5D, EIF1AY and TXLNGY). Additionally, we found that nearly half of these 21 edges involved a protein of the Metallothionein family (i.e. MT1H, MT2A, MT1HL1, MT1X and MT1G), involved in the regulation of transcription factors and in cancers (54).

The most common edges in the disease co-expression networks were then compared to the normal co-expression network to identify correlations between the two, assuming that proteins involved in these edges would have basal levels of expression and that they may not be relevant to a disease-specific context. We perform a range of comparisons on the most common edges by focusing on only the top 1000 to the top 10 000, in intervals of 1000. In order to maintain a balance in these comparisons, given the high number of edges in the normal network, we subset the edges of the normal network to an equal number of edges that is currently being compared. To do so, we sort the edges of the normal co-expression network by strongest correlations (i.e. highest absolute value of weight) and select subsets of edges from the top of the list for comparison. When the most common edges in the disease co-expression networks were compared to a proportionate range of edges with the strongest correlations in the normal network (i.e. from 1000 to 10 000 edges), we found that between 19% and 17% of the edges consistently overlapped, respectively. Focusing on these ~19% of edges that were shared between the normal and most common disease networks, we were then interested in investigating whether these edges could also be found in the interactome, STRING network, and HIPPIE network. In the interactome, we found an overlap of only 8%, with this number decreasing to 4% as the number of edges being compared in disease against normal co-expression networks increased (i.e. between the top 1,000 and 10,000 most common edges). With STRING, its overlap with edges shared between the disease and normal networks was 30%, increasing to 45%, and in HIPPIE, the overlap was consistently 8%. These findings are expected because the overlap is proportional to network size (i.e. the STRING network has 20 times more edges than the interactome while HIPPIE has twice as many). Additionally, from these 8% to 4% of edges which overlapped with the interactome, we looked at the top 10 most connected proteins, consistently identifying the same proteins as the number of edges in the comparison increased. Furthermore, we found that the direct overlap between the top 1000 most common edges of the disease networks with the interactome was only 4%, 57% with the STRING network, and 6% with the HIPPIE network; while the overlap between just the top 1000 most common edges of the disease networks which were not among the top edges of the normal network with the interactome was 2%, with the STRING network 54%, and with the HIPPIE network 5%. Because this latter group of edges represents the top edges of the disease co-expression networks (but not of

the normal) which overlap with the interactome and other PPI networks, they may also warrant further investigation as they are more likely to consistently appear across diseases than in normal networks.

Overlaying co-expression networks with pathway knowledge supports the identification of disease associated pathways

In this subsection, we systematically overlaid pathway knowledge with disease co-expression networks to reveal the consensus and/or differences between the latter group of networks and well-established protein-protein interactions in pathway databases. Given that strongly co-expressed genes can be used as a proxy for functional similarity (22), it can be inferred that genes that are co-expressed could also be involved in the same pathway. In other words, we assume that if a given pathway is relevant to a disease, the proteins in the pathway would be strongly correlated in the disease co-expression network. Thus, following this assumption, we were interested in identifying the pathways associated with each of the investigated diseases. Using pathways from KEGG, we applied two methods which, i) map pathway knowledge to disease co-expression networks and ii) map pathway knowledge to the interactome, and the mapped portion of the interactome to disease co-expression networks (see Methods).

As expected, we noted that the results of both methods were nearly identical, indicating that pathway proteins were readily mappable to the interactome. Nonetheless, we found that the second method resulted in generally higher similarity values as it only considered edges that were identifiable in the interactome, rather than edges resulting from all possible combinations of pathway proteins (Supplementary Figure S16). Overall, clearly noticeable patterns were discernible, with groups of pathways showing variable levels of similarity in specific diseases and disease clusters (Figure 6).

In particular, we observed multiple diseases/disease clusters with higher similarity values for pathways relevant to the given disease/cluster. Among these clusters, a large group of pathways showed a high degree of similarity to cognitive disorders (Figure 6; teal), including pathways for long-term potentiation, multiple neurotransmitter systems (i.e. serotonergic synapse, glutamatergic synapse, and dopaminergic synapse), long-term depression, alcoholism, and pathways for addictions (i.e. nicotine addiction, amphetamine addiction, morphine addiction, and cocaine addiction) (Supplementary Table S6). Not surprisingly, the pathway for long-term depression showed the highest similarity with the co-expression network for mental depression. Furthermore, the gastrointestinal system disease cluster (Figure 6; blue) contained co-expression networks with the highest level of similarity with several pathways, e.g. the pathways responsible for renal cell carcinoma, colorectal cancer, pathogenic *Escherichia coli* infection, intestinal immune network for IgA production, and inflammatory bowel disease (Supplementary Table S7). Additionally, a broad group of pathways showed the highest similarity values for the two reproductive system diseases (i.e. endometriosis and ovarian cancer) (Figure 6; purple) over all other diseases and disease clusters (Supplementary Table S8). Interest-

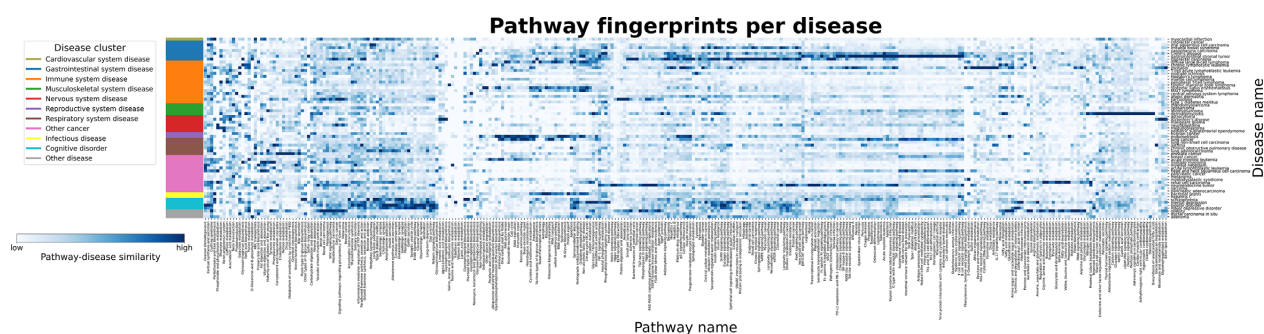


Figure 6. Mapping disease-specific expression patterns with pathway knowledge via network similarity. The heatmap illustrates the consensus similarity between KEGG pathways and disease co-expression networks. Similarity was defined as the percent of neighbors existing in a disease co-expression network out of all possible pairs of proteins from KEGG pathways (i.e. pathway-disease similarity), with lighter values corresponding to a lower similarity and darker values corresponding to higher similarity. The values (given as the percent of neighbors found) were standardized to a feature range from 0–1 for each pathway and pathways with similar values were grouped together. To ease the identification of patterns of pathway fingerprints across similar diseases, diseases were grouped by the previously defined clusters (Figure 4). A high quality version of this figure is available at <https://github.com/CoXPath/CoXPath/tree/main/results/figures>.

ingly, we found that several cancers, including gastrointestinal stromal tumor, lung cancer, head and neck squamous cell carcinoma, neuroendocrine tumor, hepatitis C, breast cancer, and ductal carcinoma in situ shared a common pattern of similar pathways (Supplementary Table S9). Among the diseases, dermatomyositis was particularly distinguishable above all others, displaying notably higher similarity to several pathways (Supplementary Table S10).

Altogether, we have demonstrated how by overlapping pathway knowledge to disease-specific co-expression networks, we can identify pathways associated with a particular disease. Additionally, we have also shown how this approach can be used to cluster diseases by the pathways they have in common, pointing to sets of potentially shared mechanisms across diseases.

Case scenario: in-depth investigation of the long-term potentiation pathway in the context of schizophrenia

In the previous section, we identified disease-associated pathways by calculating similarity between pathway knowledge and disease co-expression networks. To understand the mechanisms that underlie the similarity of a pathway to a given disease, in a case scenario, we next investigated the long-term potentiation (LTP) pathway which had yielded high similarity to the schizophrenia co-expression network. An association between this pathway and schizophrenia has already been reported in the literature, with evidence indicating impairment of LTP in the disorder (55,56).

The LTP pathway is categorized as a nervous system pathway in KEGG, with 35 edges between a set of 25 proteins/protein complexes (Figure 7). As 19 of the nodes are protein complexes containing multiple proteins, the pathway covers a total of 67 unique proteins. By overlaying the co-expression network for schizophrenia with this pathway, we identified four major edges in common, all of which were well-established interactions within this particular pathway and formed a subgraph. These edges were among the most essential of the LTP pathway; interactions between protein kinase A and the NMDA receptor, Ca^{2+} /calmodulin-dependent protein kinase II (CAMKII)

and calmodulin, and the subsequent activation of AMPAR (57) and metabotropic glutamate receptors (58) by CAMKII play key roles in determining the strength of synaptic transmission and ultimately the expression of LTP (59).

Interestingly, by overlaying the schizophrenia co-expression network with the LTP pathway, we found 53 unique correlations between proteins of the LTP pathway, indicating that the vast majority of proteins in this pathway were correlated in the co-expression network (Figure 7; grey edge), and demonstrating that indeed, proteins that are correlated in a given co-expression network can also be involved in the same biological process (31). 19 of these correlations were between calcium voltage channel complexes or calmodulin, which both have roles in the initial activation of the pathway, and other proteins (e.g. glutamate receptors). Similarly, there were approximately 20 correlations between all glutamate receptors present in the pathway and other proteins. The remaining correlations involved Erk/MAP kinase and cAMP, which ultimately regulate EP300 and CREBBP (which form the CREB binding protein complex) as well as ATF4. ATF4 is a transcription factor with multiple regulatory functions and whose polymorphisms have been associated with schizophrenia in male patients (60).

Lastly, we attempted to pinpoint candidate downstream pathways of LTP in the context of schizophrenia by investigating the edges of ATF4 given its role as a key regulator of the LTP pathway (61). As ATF4 is strongly correlated with 70 other proteins in the co-expression networks, we conducted a pathway enrichment analysis as a proxy to reveal pathway crosstalks mediated by this protein (see Methods). This analysis pinpointed four pathways from which three were involved in protein and RNA processing (i.e. ubiquitin mediated proteolysis, RNA transport, spliceosome), biological processes which have been linked with schizophrenia (62,63), while the fourth pathway, cell cycle, has also been associated with the disease (64,65) (Supplementary Table S11). These findings indicate that there may be crosstalk between these pathways that could be explored in the future.

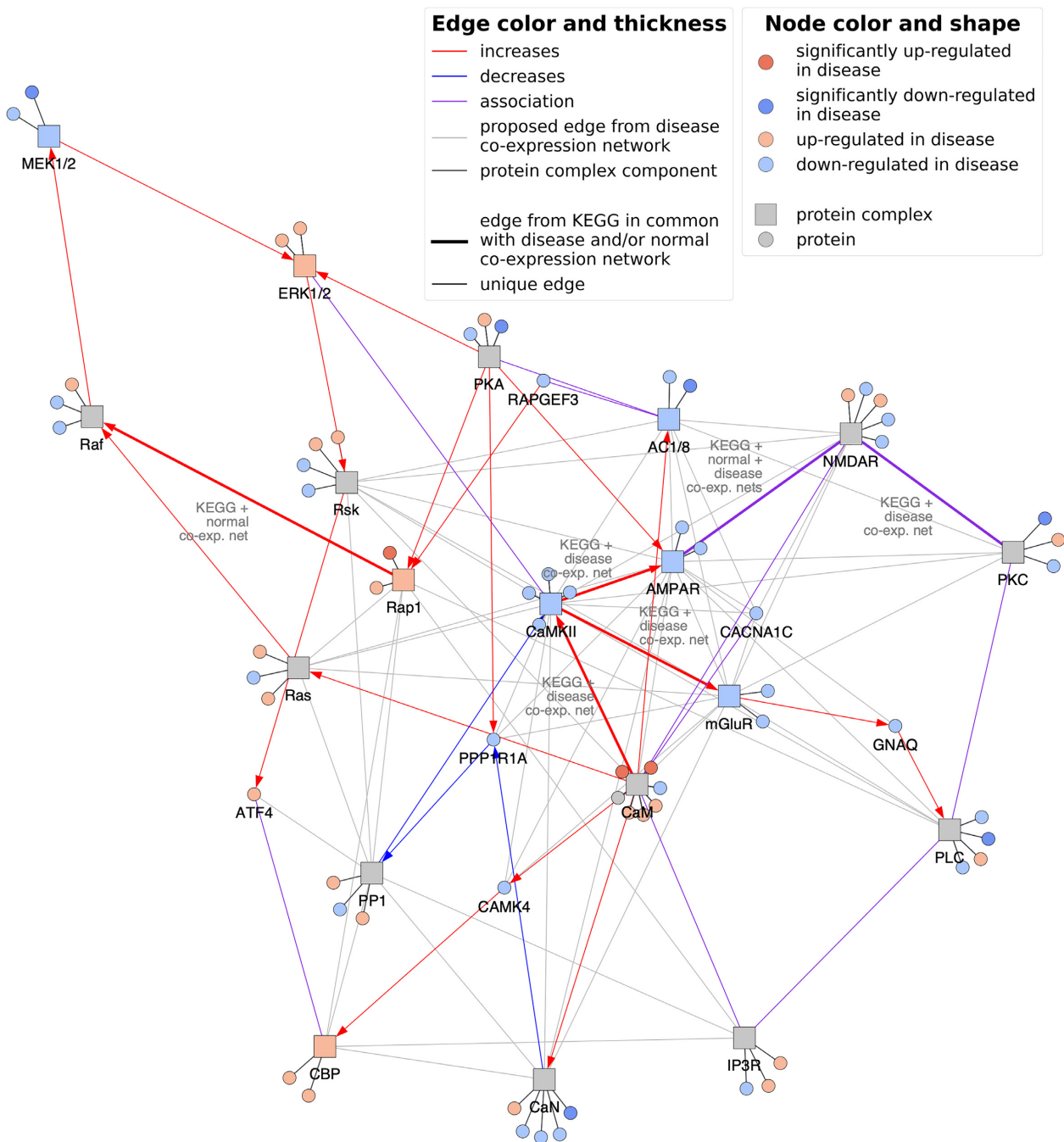


Figure 7. Long-term potentiation (LTP) pathway in the context of schizophrenia. The figure depicts the overlap of the LTP pathway with the schizophrenia co-expression network in addition to the normal co-expression network. Protein-protein interactions and associations between proteins and/or protein complexes are displayed as colored edges, while black edges denote membership of proteins to protein complexes. Edges that were common to both the LTP pathway, the disease co-expression network and/or the normal co-expression network are bolded, while grey edges denote correlations exclusively from the schizophrenia co-expression network. Differential gene expression analysis was performed and genes that were up- and down-regulated are colored orange and blue, respectively, with those that were significantly differentially expressed (i.e. adj. P value < 0.05) given less transparency. Protein complex nodes are then additionally colored if all members are in agreement with the direction of regulation. The code to generate this figure for any combination of disease co-expression network and pathway can be found at https://github.com/CoXPath/CoXPath/blob/main/analysis/3.5_analysis.ipynb.

DISCUSSION

Here, we have presented a systematic network-based approach that builds a bridge between disease signatures and pathway knowledge to better understand human pathophysiology. Our analysis has enabled us to globally evaluate the consensus between disease-specific transcriptomic data and an integrative human interactome network. Leveraging hundreds of transcriptomic datasets from over 60 major indications, we have explored the expression patterns observed in their corresponding co-expression networks at three different scales (i.e. at the node, edge and pathway levels). At each of these scales, we have investigated which proteins, subgraphs, and pathways could be associated with both disease-specific and shared mechanisms. Finally, we have presented a case scenario where we demonstrated how our approach can be used to investigate the role of a specific pathway in a disease-specific context.

There exist several limitations to this study. Firstly, we sought to improve the quality of the data by systematically integrating transcriptomic datasets from the same disease group, however, in doing so, we assumed that these datasets were equivalent. Although we attempted to address this assumption by enforcing a conservative inclusion and exclusion criteria as well as extensively curating the metadata associated with each dataset to group datasets into distinct diseases, disease heterogeneity for patients cannot be ignored. Secondly, we restricted this study to the most used platform in ArrayExpress in order to avoid possible effects caused by the array type, thus limiting the number of datasets that could potentially be used. Thirdly, since the cut-off chosen to generate the co-expression networks influences the resulting network (36), we exclusively focused on the 1% strongest correlations. While this cut-off was well-suited for our large-scale approach, in the future, less restrictive cut-offs could be used to generate co-expression networks as well as other methods. For instance, Pardo-Diaz *et al.* (66) recently presented a novel method that adds directionality into the co-expression network. Finally, while we constructed a human interactome network from multiple pathway and interaction databases, the majority of proteins from the co-expression networks could not be mapped to the network, highlighting the incompleteness of the current interactome.

Although we have demonstrated a proof-of-concept of our methodology across hundreds of datasets and in over sixty indications, we were only able to scratch the surface of the possible analyses that could be conducted with the resources generated within the context of this work. Thus, we have made the datasets and scripts generated in this study public to allow other researchers to conduct additional analyses on them. In the following, we outline several future applications and extensions of this work. Firstly, while we employed data from microarray technologies, the presented analysis could be expanded and/or validated by incorporating datasets generated from other platforms and technologies (e.g. RNASeq) or deposited in other databases such as GEO (28) which, in turn, can facilitate the discovery of novel genes as well as allow us to add new indications and validate the current mechanisms identified in our analysis, respectively. However, conducting such an analy-

sis would require extensive harmonization efforts at both the data and metadata level given the differences across chips and technologies, and the lack of structured metadata present in transcriptomic experiments. Secondly, the disease-specific co-expression networks generated in this work could be compared against well-established databases such as DisGeNet (67) and OMIM (68) to propose novel gene-disease associations that can be integrated into these resources. Thirdly, other advanced network analysis methods could be conducted to analyze specific network motifs in the future. Fourthly, with prior enrichment of the presented networks with drug-target information, network-based drug discovery methods can be applied to identify candidate drugs and druggable pathways for the particular disease condition(s) (69–72). Finally, another potential line of research would be to apply our methodology on datasets generated from a variety of cell lines to identify cell-specific transcriptional patterns.

DATA AVAILABILITY

All data supporting the conclusions of this article are available at <https://doi.org/10.5281/zenodo.4700652> and all scripts can be found at <https://github.com/CoXPath/CoXPath>.

SUPPLEMENTARY DATA

[Supplementary Data](#) are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank Sumit Madan for his help running the ZOOMA queries, Carlos Bobis-Álvarez for his assistance grouping similar indications, Chris W. Diana, Helena Hermanowski and Carina Steinborn for their assistance generating the figures, and Daniel Rivas-Barragan for his valuable feedback.

Authors contributions: M.H.A. conceived the original idea. D.D.F. designed and supervised the study with assistance from SM. T.R. implemented the pipeline to download, process and categorize the gene expression datasets. R.Q.F. and T.R. generated the co-expression networks for each group. S.M. and D.D.F. generated the interactome network. R.Q.F. performed the analyses. R.Q.F., T.R., S.M. and D.D.F. interpreted the results. A.T.K., M.H.A. and D.D.F. acquired the funding. R.Q.F., T.R., S.M. and D.D.F. wrote the manuscript.

All authors have read and approved the final manuscript.

FUNDING

Fraunhofer Cluster of Excellence ‘Cognitive Internet Technologies’; German Federal Ministry of Education and Research (BMBF) [01ZX1904C]. Funding for open access charge: Institute for Algorithms and Scientific Computing. *Conflict of interest statement.* D.D.F. received salary from Enveda Biosciences.

REFERENCES

- UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, **47**, D506–D515.
- Caldera, M., Buphamalai, P., Müller, F. and Menche, J. (2017) Interactome-based approaches to human disease. *Curr. Opin. Syst. Biol.*, **3**, 88–94.
- Franzese, N., Groce, A., Murali, T.M. and Ritz, A. (2019) Hypergraph-based connectivity measures for signaling pathway topologies. *PLoS Comput. Biol.*, **15**, e1007384.
- Winterbach, W., VanMieghem, P., Reinders, M., Wang, H. and deRidder, D. (2013) Topology of molecular interaction networks. *BMC Syst. Biol.*, **7**, 90.
- Hanspers, K., Riutta, A., Summer-Kutmon, M. and Pico, A.R. (2020) Pathway information extracted from 25 years of pathway figures. *Genome Biol.*, **21**, 273.
- Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M. and Tanabe, M. (2021) KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res.*, **49**, D545–D551.
- Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., Sidiropoulos, K., Cook, J., Gillespie, M., Haw, R. *et al.* (2020) The reactome pathway knowledgebase. *Nucleic Acids Res.*, **48**, D498–D503.
- Reimand, J., Isserlin, R., Voisin, V., Kucera, M., Tannus-Lopes, C., Rostamianfar, A., Wadi, L., Meyer, M., Wong, J., Xu, C. *et al.* (2019) Pathway enrichment analysis and visualization of omics data using G:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nat. Protoc.*, **14**, 482–517.
- Segura-Lepe, M.P., Keun, H.C. and Ebbels, T.M. (2019) Predictive modelling using pathway scores: robustness and significance of pathway collections. *BMC Bioinformatics*, **20**, 543.
- Huang, J.K., Carlin, D.E., Yu, M.K., Zhang, W., Kreisberg, J.F., Tamayo, P. and Ideker, T. (2018) Systematic evaluation of molecular networks for discovery of disease genes. *Cell Syst.*, **6**, 484–495.
- Mubeen, S., Hoyt, C.T., Gemünd, A., Hofmann-Apitius, M., Fröhlich, H. and Domingo-Fernández, D. (2019) The impact of pathway database choice on statistical enrichment analysis and predictive modeling. *Frontiers in Genetics*, **10**, 1203.
- vanDam, S., Vösa, U., vanderGraaf, A., Franke, L. and deMagalhães, J.P. (2018) Gene co-expression analysis for functional classification and gene–disease predictions. *Brief. Bioinform.*, **19**, 575–592.
- Langfelder, P. and Horvath, S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, **9**, 559.
- Paci, P., Colombo, T., Fiscon, G., Gurtner, A., Pavesi, G. and Farina, L. (2017) SWIM: a computational tool to unveiling crucial nodes in complex biological networks. *Sci. Rep.*, **7**, 44797.
- Margolin, A.A., Nemenman, L., Basso, K., Wiggins, C., Stolovitzky, G., DallaFavera, R. and Califano, A. (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, **7**, S7.
- Stuart, J.M., Segal, E., Koller, D. and Kim, S.K. (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science*, **302**, 249–255.
- Chou, W.C., Cheng, A.L., Brotto, M. and Chuang, C.Y. (2014) Visual gene-network analysis reveals the cancer gene co-expression in human endometrial cancer. *BMC Genomics*, **15**, 300.
- Xiang, S., Huang, Z., Wang, T., Han, Z., Christina, Y.Y., Ni, D., Huang, K. and Zhang, J. (2018) Condition-specific gene co-expression network mining identifies key pathways and regulators in the brain tissue of Alzheimer's disease patients. *BMC Med. Genet.*, **11**, 115.
- Gene Ontology Consortium. (2019) The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res.*, **47**, D330–D338.
- Mao, Y., Nie, Q., Yang, Y. and Mao, G. (2020) Identification of co-expression modules and hub genes of retinoblastoma via co-expression analysis and protein-protein interaction networks. *Mol. Med. Rep.*, **22**, 1155–1168.
- Yao, Q., Zhenyu, S., Wang, B. and Qin, Q. (2019) Identifying key genes and functionally enriched pathways in sjögren's syndrome by weighted gene Co-Expression network analysis. *Front. Genet.*, **10**, 1142.
- Paci, P., Fiscon, G., Conte, F., Wang, R.S., Lorenzo, F. and Loscalzo, J. (2021) Gene co-expression in the interactome: moving from correlation toward causation via an integrated approach to disease module discovery. *NPJ Syst. Biol. Appl.*, **7**, 3.
- Falcone, R., Conte, F., Fiscon, G., Pecce, V., Sponziello, M., Durante, C., Farina, L., Filetti, S., Paci, P. and Verrienti, A. (2019) BRAF V600E-mutant cancers display a variety of networks by SWIM analysis: prediction of vemurafenib clinical response. *Endocrine*, **64**, 406–413.
- Fiscon, G., Conte, F., Licursi, V., Nasi, S. and Paci, P. (2018) Computational identification of specific genes for glioblastoma stem-like cells identity. *Sci. Rep.*, **8**, 7769.
- Fiscon, G., Conte, F. and Paci, P. (2018) SWIM tool application to expression data of glioblastoma stem-like cell lines, corresponding primary tumors and conventional glioma cell lines. *BMC Bioinformatics*, **19**, 103–121.
- Paci, P., Fiscon, G., Conte, F., Licursi, V., Morrow, J., Hersh, C., Cho, M., Castaldi, P., Glass, K. and Silverman, E.K. (2020) Integrated transcriptomic correlation network analysis identifies COPD molecular determinants. *Sci. Rep.*, **10**, 3361.
- Athar, A., Füllgrabe, A., George, N., Iqbal, H., Huerta, L., Ali, A., Snow, C., Fonseca, N.A., Petryszak, R., Papatheodorou, I. *et al.* (2019) ArrayExpress update—from bulk to single-cell expression data. *Nucleic Acids Res.*, **47**, D711–D715.
- Edgar, R., Domrachev, M. and Lash, A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.
- Obayashi, T., Kagaya, Y., Aoki, Y., Tadaka, S. and Kinoshita, K. (2019) COXPRESdb v7: a gene coexpression database for 11 animal species supported by 23 coexpression platforms for technical evaluation and evolutionary inference. *Nucleic Acids Res.*, **47**, D55–D62.
- Doncheva, N.T., Palasca, O., Yarani, R., Litman, T., Anthon, C., Groenen, M.A.M., Stadler, P.F., Pociot, F., Jensen, L.J. and Gorodkin, J. (2021) Human pathways in animal models: possibilities and limitations. *Nucleic Acids Res.*, **49**, 1859–1871.
- Vella, D., Zoppis, I., Mauri, G., Mauri, P. and DiSilvestre, D. (2017) From protein-protein interactions to protein co-expression networks: a new perspective to evaluate large-scale proteomic data. *EURASIP J. Bioinformatics Syst. Biol.*, **2017**, 6.
- Schriml, L.M., Mitraka, E., Munro, J., Tauber, B., Schor, M., Nickle, L., Felix, V., Jeng, L., Bearer, C., Lichenstein, R. *et al.* (2018) Human Disease Ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Res.*, **47**, D955–D962.
- Johnson, W.E., Li, C. and Rabinovic, A. (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, **8**, 118–127.
- Allen, J.D., Xie, Y., Chen, M., Girard, L. and Xiao, G. (2012) Comparing statistical methods for constructing large scale gene networks. *PLoS One*, **7**, e29348.
- Perkins, A.D. and Langston, M.A. (2009) Threshold selection in gene co-expression networks using spectral graph theory techniques. *BMC Bioinformatics*, **10**, S4.
- Yip, A.M. and Horvath, S. (2007) Gene network interconnectedness and the generalized topological overlap measure. *BMC Bioinformatics*, **8**, 22.
- Martens, M., Ammar, A., Riutta, A., Waagmeester, A., Sletter, D.N., Hanspers, K., Miller, R.A., Digles, D., Lopes, E.N., Ehrhart, F. *et al.* (2020) WikiPathways: connecting communities. *Nucleic Acids Res.*, **49**, D613–D621.
- Oughtred, R., Stark, C., Breitkreutz, B.J., Rust, J., Boucher, L., Chang, C., Kolas, N., O'Donnell, L., Leung, G., McAdam, R. *et al.* (2019) The BioGRID interaction database: 2019 update. *Nucleic Acids Res.*, **47**, D529–D541.
- Orchard, S., Ammari, M., Aranda, B., Breuzza, L., Briganti, L., Broackes-Carter, F., Campbell, N.H., Chavali, G., Chen, C., del-Toro, N. *et al.* (2014) The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.*, **42**, D358–D363.
- Rodchenkov, I., Babur, O., Luna, A., Aksoy, B.A., Wong, J.V., Fong, D., Franz, M., Siper, M.C., Cheung, M., Wrana, M. *et al.* (2020) Pathway Commons 2019 Update: integration, analysis and exploration of pathway data. *Nucleic Acids Res.*, **48**, D489–D497.
- Domingo-Fernández, D., Mubeen, S., Marin-Llao, J., Hoyt, C. and Hofmann-Apitius, M. (2019) PathMe: Merging and exploring mechanistic pathway knowledge. *BMC Bioinformatics*, **20**, 243.

42. Alanis-Lobato, G., Andrade-Navarro, M.A. and Schaefer, M.H. (2016) HIPPIE v2. 0: enhancing meaningfulness and reliability of protein–protein interaction networks. *Nucleic Acids Res.*, **45**, D408–D414.
43. Szklarczyk, D., Gable, A., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H., Bork, P. *et al.* (2019) STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.*, **47**, D607–D613.
44. Hagberg, A., Swart, P. and Chult, D.S. (2008) Exploring network structure, dynamics, and function using NetworkX. In: Varoquaux, G., Vaught, T. and Millman, J. (eds). *Proceedings of the 7th Python in Science conference (SciPy 2008)*. pp. 11–15.
45. Smyth, G.K. (2005) Limma: linear models for microarray data. In: *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. pp. 397–420.
46. Fisher, R.A. (1992) Statistical methods for research workers. In: *Breakthroughs in Statistics*. Springer, NY, pp. 66–70
47. Benjamini, Y. and Yekutieli, D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.*, **29**, 1165–1188.
48. Cassandri, M., Smirnov, A., Novelli, F., Pitolli, C., Agostini, M., Malewicz, M., Melino, G. and Raschella, G. (2017) Zinc-finger proteins in health and disease. *Cell Death Discov.*, **3**, 17071.
49. Gort, E.H., VanHaften, G., Verlaan, I., Groot, A.J., Plasterk, R.H.A., Shvarts, A., Suijkerbuijk, K.P.M., van Laar, T., van der Wall, E., Raman, V. *et al.* (2008) The TWIST1 oncogene is a direct target of hypoxia-inducible factor-2 α . *Oncogene*, **27**, 1501–1510.
50. Peterson, T.R., Laplante, M., Thoreen, C.C., Sancak, Y., Kang, S.A., Kuehl, W.M., Gray, N.S. and Sabatini, D.M. (2009) DEPTOR is an mTOR inhibitor frequently overexpressed in multiple myeloma cells and required for their survival. *Cell*, **137**, 873–886.
51. Ratsada, P., Hijiy, N., Hidano, S., Tsukamoto, Y., Nakada, C., Uchida, T., Kobayashi, T. and Moriyama, M. (2020) DUSP4 is involved in the enhanced proliferation and survival of DUSP4-overexpressing cancer cells. *Biochem. Biophys. Res. Commun.*, **528**, 586–593.
52. Schmid, C.A., Robinson, M.D., Scheifinger, N.A., Müller, S., Cogliatti, S., Tzankov, A. and Müller, A. (2015) DUSP4 deficiency caused by promoter hypermethylation drives JNK signaling and tumor cell survival in diffuse large B cell lymphoma. *J. Exp. Med.*, **212**, 775–792.
53. Spencer, S.A., Suárez-Pozos, E., Escalante, M., Myo, Y.P. and Fuss, B. (2020) Sodium–calcium exchangers of the SLC8 family in oligodendrocytes: functional properties in health and disease. *Neurochem. Res.*, **45**, 1287–1297.
54. Gumulec, J., Raudenska, M., Adam, V., Kizek, R. and Masarik, M. (2014) Metallothionein–immunohistochemical cancer biomarker: a meta-analysis. *PLoS One*, **9**, e85346.
55. Frantseva, M.V., Fitzgerald, P.B., Chen, R., Möller, B., Daigle, M. and Daskalakis, Z.J. (2008) Evidence for impaired long-term potentiation in schizophrenia and its relationship to motor skill learning. *Cereb. Cortex*, **18**, 990–996.
56. Hasan, A., Nitsche, M.A., Rein, B., Schneider-Axmann, T., Guse, B., Gruber, O., Falkai, P. and Wobrock, T. (2011) Dysfunctional long-term potentiation-like plasticity in schizophrenia revealed by transcranial direct current stimulation. *Behav. Brain Res.*, **224**, 15–22.
57. Kristensen, A.S., Jenkins, M.A., Banke, T.G., Schousboe, A., Makino, Y., Johnson, R.C., Huganir, R. and Traynelis, S.F. (2011) Mechanism of Ca²⁺/calmodulin-dependent kinase II regulation of AMPA receptor gating. *Nat. Neurosci.*, **14**, 727–735.
58. Foster, W.J., Taylor, H.B., Padamsey, Z., Jeans, A.F., Galione, A. and Emptage, N.J. (2018) Hippocampal mGluR1-dependent long-term potentiation requires NAADP-mediated acidic store Ca²⁺ signaling. *Sci. Signal*, **11**, eaat9093.
59. Herring, B.E. and Nicoll, R.A. (2016) Long-term potentiation: from CaMKII to AMPA receptor trafficking. *Annu. Rev. Physiol.*, **78**, 351–365.
60. Qu, M., Tang, F., Wang, L., Yan, H., Han, Y., Yan, J., Yue, W. and Zhang, D. (2008) Associations of ATF4 gene polymorphisms with schizophrenia in male patients. *Am. J. Med. Genet. Part B: Neuropsychiatric Genet.*, **147**, 732–736.
61. Pasini, S., Corona, C., Liu, J., Greene, L.A. and Shelanski, M.L. (2015) Specific downregulation of hippocampal ATF4 reveals a necessary role in synaptic plasticity and memory. *Cell Rep.*, **11**, 183–191.
62. McInnes, L.A. and Lauriat, T.L. (2006) RNA metabolism and dysmyelination in schizophrenia. *Neurosci. Biobehav. Rev.*, **30**, 551–561.
63. Glatt, S.J., Cohen, O.S., Faraone, S.V. and Tsuang, M.T. (2011) Dysfunctional gene splicing as a potential contributor to neuropsychiatric disorders. *Am. J. Med. Genet. Part B: Neuropsychiatric Genet.*, **156**, 382–392.
64. Fan, Y., Abrahamsen, G., McGrath, J.J. and Mackay-Sim, A. (2012) Altered cell cycle dynamics in schizophrenia. *Biol. Psychiatry*, **71**, 129–135.
65. Katsel, P., Davis, K.L., Li, C., Tan, W., Greenstein, E., Hoffman, L.B.K. and Haroutunian, V. (2008) Abnormal indices of cell cycle activity in schizophrenia and their potential association with oligodendrocytes. *Neuropsychopharmacology*, **33**, 2993–3009.
66. Pardo-Diaz, J., Bozhilova, L.V., Beguerisse-Diaz, M., Poole, P.S., Deane, C.M. and Reinert, G. (2021) Robust gene coexpression networks using signed distance correlation. *Bioinformatics*, btab041.
67. Piñero, J., Bravo, A., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., García-García, J., Sanz, F., Furlong, L.I. *et al.* (2016) DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.*, **45**, D833–D839.
68. Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A. and McKusick, V.A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **33**, D514–D517.
69. Peyvandipour, A., Saberian, N., Shafi, A., Donato, M. and Draghici, S. (2018) A novel computational approach for drug repurposing using systems biology. *Bioinformatics*, **34**, 2817–2825.
70. Rivas-Barragan, D., Mubeen, S., Bernat, F.G., Hofmann-Apitius, M. and Domingo-Fernández, D. (2020) Drug2ways: reasoning over causal paths in biological networks for drug discovery. *PLoS Comput. Biol.*, **16**, e1008464.
71. Fiscon, G. and Paci, P. (2021) SAveRUNNER: an R-based tool for drug repurposing. *BMC Bioinformatics*, **22**, 150.
72. Fiscon, G., Conte, F., Farina, L. and Paci, P. (2021) SAveRUNNER: a network-based algorithm for drug repurposing and its application to COVID-19. *PLoS Comput. Biol.*, **17**, e1008686.

A.5 Elucidating gene expression patterns across multiple biological contexts through a large-scale investigation of transcriptomic datasets

Reprinted with permission from “Figueiredo R.Q., Díaz del Ser S., Raschka T., Hofmann-Apitius M., Kodamullil A.T., Mubeen S.[†], and Domingo-Fernández D.[†] (2022). Elucidating gene expression patterns across multiple biological contexts through a large-scale investigation of transcriptomic datasets. *BMC Bioinformatics*, 23(1)”.

Copyright © Figueiredo R.Q., *et al.*, 2022.

RESEARCH

Open Access



Elucidating gene expression patterns across multiple biological contexts through a large-scale investigation of transcriptomic datasets

Rebeca Queiroz Figueiredo^{1,2}, Sara Díaz del Ser^{1,2}, Tamara Raschka^{1,2,3}, Martin Hofmann-Apitius^{1,2}, Alpha Tom Kodamullil¹, Sarah Mubeen^{1,2,3†} and Daniel Domingo-Fernández^{1,3,4*†}

[†]Sarah Mubeen and Daniel Domingo-Fernández contributed equally to this work

*Correspondence: daniel.domingo.fernandez@scai.fraunhofer.de

¹ Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing, 53757 Sankt Augustin, Germany

² Bonn-Aachen International Center for IT, Rheinische Friedrich-Wilhelms-Universität Bonn, 53115 Bonn, Germany

³ Fraunhofer Center for Machine Learning, Sankt Augustin, Germany

⁴ Enveda Biosciences, Boulder, CO 80301, USA

Abstract

Distinct gene expression patterns within cells are foundational for the diversity of functions and unique characteristics observed in specific contexts, such as human tissues and cell types. Though some biological processes commonly occur across contexts, by harnessing the vast amounts of available gene expression data, we can decipher the processes that are unique to a specific context. Therefore, with the goal of developing a portrait of context-specific patterns to better elucidate how they govern distinct biological processes, this work presents a large-scale exploration of transcriptomic signatures across three different contexts (i.e., tissues, cell types, and cell lines) by leveraging over 600 gene expression datasets categorized into 98 subcontexts. The strongest pairwise correlations between genes from these subcontexts are used for the construction of co-expression networks. Using a network-based approach, we then pinpoint patterns that are unique and common across these subcontexts. First, we focused on patterns at the level of individual nodes and evaluated their functional roles using a human protein–protein interactome as a referential network. Next, within each context, we systematically overlaid the co-expression networks to identify specific and shared correlations as well as relations already described in scientific literature. Additionally, in a pathway-level analysis, we overlaid node and edge sets from co-expression networks against pathway knowledge to identify biological processes that are related to specific subcontexts or groups of them. Finally, we have released our data and scripts at <https://zenodo.org/record/5831786> and <https://github.com/ContNeXt/>, respectively and developed ContNeXt (<https://contnext.scai.fraunhofer.de/>), a web application to explore the networks generated in this work.

Keywords: Transcriptomic, Biological context, Co-expression networks, Gene expression, Network biology



© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Introduction

While gene expression profiling has markedly improved our understanding of the molecular underpinnings of biological processes, the knowledge we acquire from a particular study performed within a given context may not generalize to another. For instance, accumulating evidence shows that average gene expression varies extensively across cell lines or tissues of the same organism [38, 43] as well as across species [32]. Context-specificity has also been noted when investigating the reproducibility of protein–protein interactions (PPIs) across conditions in literature-curated PPI databases in Stacey et al. [39], finding no evidence for the occurrence of anywhere from 19 to 55% of interactions reported in these databases. These findings, however, are not altogether surprising given that PPI databases often store interactions that occur across various experimental conditions and contexts which may fail to be observed if either of these were to vary. Crucially, it is often these context-specific differences which are responsible for the variability of functions and unique characteristics of diverse cell types and tissues and their investigation is thus fundamental in understanding human biology.

Gene expression patterns that are specific to certain cell types or tissues can help us to better understand normal human physiology (e.g., which biological processes occur in specific cell types or tissues) as well as development biology (e.g., which genes are expressed in specific cell types or tissues at various developmental stages), and several studies have investigated differences in these two contexts. Specifically, Pierson et al. [30] and Dobrin et al. [5] analyzed gene expression patterns at the tissue-level, revealing function-specific patterns and subnetworks associated with obesity. Similarly, McKenzie et al. [24] analyzed co-expression changes in different cell types of the brain, discovering significant cell type-specific expression signatures, while also finding well-known cell type marker genes among the most enriched genes across cell types.

Another relevant context is cell line information, as these are widely used for the study of biological processes. In particular, cancer cell lines, such as HeLa, are frequently employed, having had many interactions characterized on them and representing the foremost models for the study of cancer biology as well as numerous other disease and normal conditions. Nonetheless, even cell lines classified to the same tissue can exhibit significant differences in gene expression [19]. For example, a study by Yu et al. [46] found that certain cell lines may not resemble the primary cells from which they originated. The discrepancies in regulation patterns across specific cell lines deem it necessary to employ tools such as the CellExpress system (developed by Lee et al. [19] which enables the analysis of over 4000 cancer cell lines for differences in gene expression levels) and resources such as the TCGA-110-CL cell line panel [46] to identify which cell lines are more suitable for a given study.

Biological networks of different types can be used to represent patterns characteristic to a particular context. These context-specific networks can be categorized based on whether they are directly derived from knowledge or data. Rachlin et al. [31] and Stacey et al. [39] are two illustrations of knowledge-driven approaches where authors generated context-specific PPI networks by leveraging information about biological processes from GO (The Gene Ontology Consortium et al. [41]) and co-occurrence literature, respectively. Similarly, the analysis of transcriptomic data through the construction of gene co-expression networks (Langfelder et al. [18]) can also serve to better understand

context-specific patterns within datasets [28]. Finally, hybrid approaches, as demonstrated by Kitsak et al. [16], have leveraged gene expression data from 64 different tissues and mapped genes expressed in specific tissues to a protein–protein interactome, revealing that these disease context-specific genes tend to be located in close proximity within the interactome. It is important to note that while transcriptomic experiments are often used as a proxy to reflect protein expression, the correlation between the two is often below 0.5 on average [26, 40]. Nevertheless, correlations between genes whose mRNA is differentially expressed and their protein products have been shown to be significantly higher than genes whose mRNA is not differentially expressed, lending support to the use of differential mRNA expression to infer changes at the protein level [17].

One of the challenges in conducting these hybrid approaches (i.e., approaches that combine data- and knowledge- derived networks) is the limited availability of context-specific resources on a large-scale (e.g., hundreds of experiments conducted within the same or similar conditions or context-specific interactomes). While there are several co-expression databases dedicated to storing context-specific information, such as species [27] and [20], the vast majority of transcriptomic datasets are not annotated with context information and thus, cannot be systematically leveraged to conduct contextualized analyses on a large-scale. Nonetheless, the Gemma system [21] has been made available to provide thousands of curated datasets, thus, more easily enabling data reuse and secondary analyses.

In this work, we apply a network-based approach to investigate transcriptomic patterns observed in a variety of subcontexts classified under three major biological contexts (i.e., tissues, cell types, and cell lines) by leveraging over 600 gene expression datasets (Fig. 1A). To do so, we first construct co-expression networks that capture the strongest gene expression correlations observed in each subcontext (Fig. 1B). Subsequently, a series of network-based analyses are conducted to enable the exploration of the similarities and differences across co-expression networks and provide insights on gene co-expression patterns across contexts (Fig. 1C). Furthermore, we study the consensus between patterns identified in the co-expression network and a human protein–protein interactome as well as pathways knowledge. Finally, we present ContNeXt, a web application we have developed to enable researchers to explore and reuse our work.

Methodology

Gene expression datasets

We identified publicly available transcriptomic datasets from each of the three contexts evaluated (i.e., tissues, cell types, and cell lines) using Gemma, a manually curated database containing metadata for over 10,000 datasets [21, 48] (Fig. 1A). This metadata is programmatically accessible through Gemma's API (<https://gemma.msl.ubc.ca/resources/restapidocs>) and is annotated using different ontologies. Specifically, for each of the three contexts of interest, the following ontologies were used: (i) UBERON for tissues [25], (ii) Cell Ontology (CL) for cell types [4], and (iii) Cell Line Ontology (CLO) for cell lines [34].

Leveraging the metadata from Gemma, we were able to classify the samples from each dataset to their corresponding context(s). To guarantee the quality of the annotations, we conducted an additional manual curation step where we confirmed that the Gemma

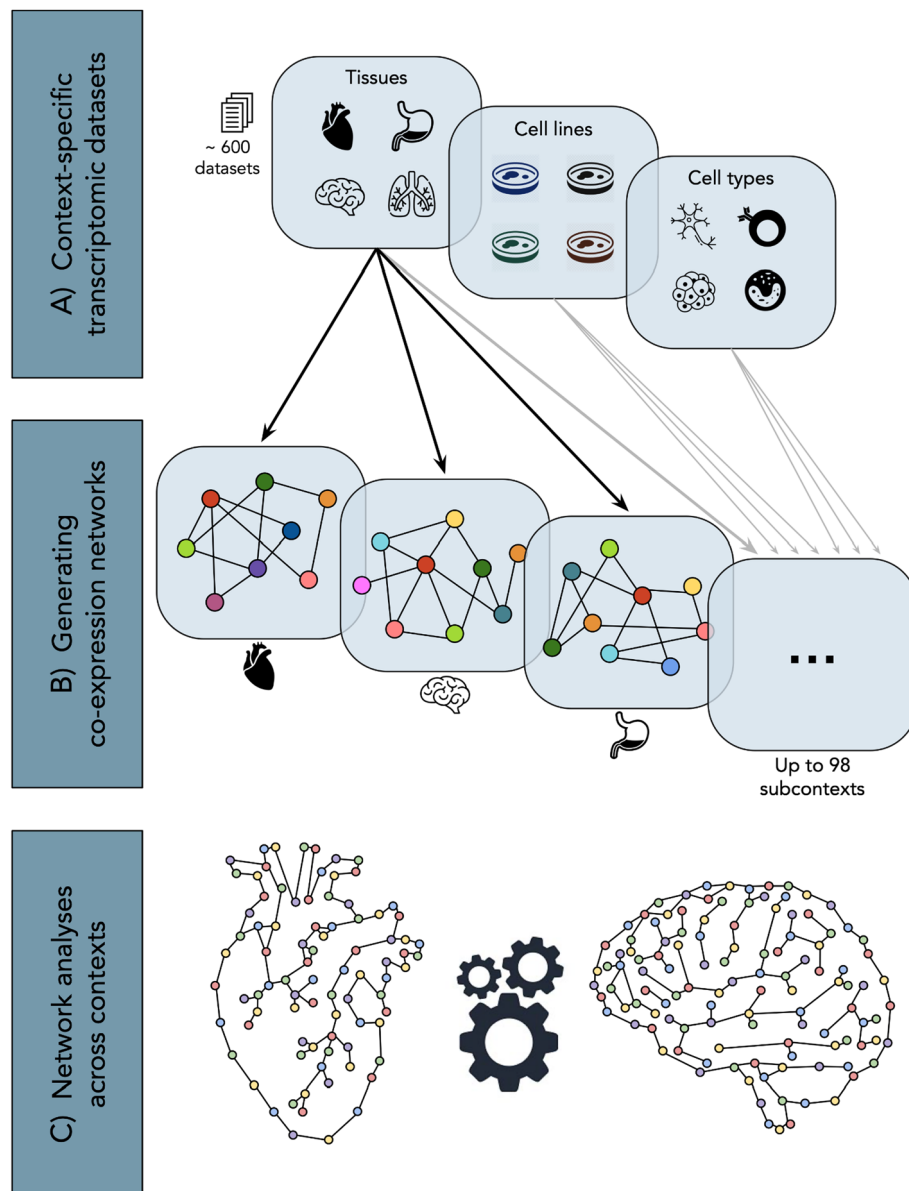


Fig. 1 Conceptualization of the presented study. **A** Over 600 context-specific transcriptomic datasets are collected and classified into 98 subcontexts (e.g., heart, astrocyte, and HeLa cell) under 3 major contexts (i.e., tissues, cell types, and cell lines), leveraging the Gemma database [21, 48] **B** Co-expression networks comprising the most strongly correlated edges observed in each subcontext are generated. **C** Network analyses provide insights on both common and unique patterns across the multiple contexts studied

sample annotations matched an ontology term for the given context present in the meta-data, if available. Additionally, we filtered out samples that were not control or reference samples as our work focuses on comparing a normal physiological state in a variety of contexts. Finally, Gemma also includes annotations on dataset quality and samples that were annotated as unusable were excluded from our study.

After the initial annotation and curation steps, we implemented scripts for the downloading and processing of datasets found in Gene Expression Omnibus (GEO) [6]. While GEO incorporates several platforms, each measures different transcripts and requires a

dedicated pipeline, and merging data from several platforms is a complicated task which can introduce biases from probe sequences, arrays, or laboratory effects. Furthermore, conducting analyses combining raw data from multiple platforms can also introduce biases [33]. Thus, our work focuses on the most commonly used platform for humans, the Affymetrix GeneChip Human Genome U133 Plus 2.0 Array platform (accession on GEO: GPL570). Out of 10,388 datasets in Gemma as of 22/04/2021, 9778 were filtered out while 610 remained for any one of the three contexts. In total, the tissue context was divided into 46 subcontexts, while the cell line and cell type contexts each contained 22 and 30 subcontexts, respectively (see Additional file 1: Tables S1–S3).

Generating co-expression networks from gene expression data

Co-expression networks were constructed using the WGCNA package in R (Langfelder et al. [18]). We followed the same procedure outlined in our previous work [9] to define the co-expression networks (Fig. 1B). This procedure focuses on the 1% highest similarity in the topological overlap matrix (TOM) to define the co-expression network for each subcontext; thus, facilitating the comparison of networks of the same size using a conservative cut-off in benchmark studies [29]. Given the platform used in this study, the most similar 1% in the TOM corresponds to 2,036,667 edges. We would like to note that the 1% cut-off is required as otherwise the networks would be fully connected, while we intend to focus only on the edges representing the most relevant transcriptomic patterns observed within each context. As edges representing a high topological overlap are also highly correlated in the TOM, we interchangeably refer to these edges as correlations for simplicity. Although this is not precise, the TOM value is based on the signed correlation but also takes the connectedness of nodes into account.

To run WGCNA, we used the raw expression data in the form of CEL-files. Each dataset was individually pre-processed with the RMA function of the *oligo* R package to conduct background subtraction and quantile normalization. Next, we merged all samples from different datasets that belong to the same subcontext and applied batch correction using ComBat [14]. Regarding the mapping of the probes to genes, if there were multiple probes mapping to the same gene, we kept the most variable probe.

Protein—protein interaction network

We built a human protein–protein interactome as described in our previous work [9] as a knowledge template to compare against the co-expression networks generated. The interactome comprises interactions from well-established databases, including KEGG [15] and Reactome [13]. This network aims at representing the set of interactions that can occur in a physiological context, though it is worth mentioning that each of these interactions may not necessarily be occurring in a particular context at any given time.

Analyses

Controllability analysis

One of the more advanced techniques in analyzing networks is examining its controllability. We employed an algorithm developed by Liu et al. [22] which explores control theory to study the controllability of a directed network and thus identify driver nodes (i.e., the set of nodes that can offer control over the whole network) in order to classify each node and

edge in a network as indispensable, dispensable, or neutral. Ideally, minimizing the number of driver nodes offers adequate control over the network regarding the given biological system's dynamics. Using this algorithm, both nodes and edges can be classified as indispensable, dispensable, or neutral if their removal creates the need to increase, decrease, or cause no change in the number of driver nodes, respectively, so that controllability is maintained.

Pairwise co-expression network similarity

To evaluate similarity across co-expression networks, we calculated the overlap of edges across each pair of co-expression networks within a given context. Since all co-expression networks have the same number of edges, the number of shared edges between networks is readily comparable without the need to normalize values.

Similarity between co-expression networks and the interactome

We assessed the similarity of each co-expression network to the human interactome by calculating the number of shared edges. Here, it is important to note that edge directionality is ignored in the interactome since co-expression networks are inherently undirected. Furthermore, we evaluated the significance of the overlap by comparing the interactome to 1000 permuted co-expression networks. Permuted versions of the co-expression networks were created using the XSwap algorithm [12] (source code available at <https://github.com/hetio/xswap>), which ensures that the permuted versions preserve the structure of the original network (i.e., all edges are shuffled while maintaining the degree of each node).

Pathway—co-expression network similarity

To investigate the correspondence of transcriptomic signatures from co-expression networks with pathway knowledge, each of the context-specific co-expression networks were overlaid with pathways from KEGG [15]. The KEGG database was exclusively employed as it contains a feasible number of pathways for analysis (i.e., less than 350). For each gene set of a given pathway P from KEGG, we calculate every pairwise combination of nodes (C_n) in P to determine the fraction of node combination pairs in C_n that exist as an edge in a given co-expression network $N = (n', E_N)$ where n' is the set of nodes in the co-expression network and E_N is the set of edges which connect the nodes n' . We term this the edge overlap, where $edge\ overlap = |\{\forall e_{u,v} s.t. (u, v) \in C_n \wedge u, v \in n' \wedge e_{u,v} \in E_N\}|$. The proportion of C_n that is in the edge overlap is the pathway-network similarity (Eq. 1). Using the pathway-network similarity, we create a similarity matrix with each network of a given context against every pathway from KEGG. This matrix is subsequently used to create a heatmap and hierarchical clustering of the co-expression networks is performed using Euclidean distances of their similarities to pathways.

Similarity between a pathway and co-expression network.

$$pathway - network\ similarity(P, N) = \frac{edge\ overlap}{|C_n|} \quad (1)$$

Implementation

Scripts to retrieve and process the datasets as well as to deploy the web application are available at <https://github.com/ContNeXt>. We have also provided comprehensive

documentation to modify the filtering steps and add extensions to the scripts. For network analysis and visualizations, we used the Python NetworkX library [11] (<https://networkx.github.io/>), and Matplotlib, and seaborn, respectively. The processed data used in this work is available at Zenodo at <https://zenodo.org/record/5831786>.

Results

In “Overview of co-expression networks and interactome” section, we provide an overview of the co-expression and PPI networks, while in “Analyses at the protein-level”, “Analyses at the network-level” and “Mapping co-expression networks to pathway knowledge” sections, we outline each of the analyses conducted, specifically at the protein-, network-, and pathway- levels (Fig. 2). Finally, “ContNeXt—a web application to explore gene expression patterns across contexts” section presents ContNeXt, a web application developed to explore the results of this work.

Overview of co-expression networks and interactome

From 364, 222, and 103 (at times overlapping) datasets that were categorized into 46 distinct tissues, 30 distinct cell types, 22 distinct cell lines, respectively, we systematically constructed co-expression networks corresponding to each of these contexts. The exact breakdown of the number of datasets and samples for each subcontext can be found in Additional file 1: Tables S1–S3. Figure 3 summarizes the size of each corresponding co-expression network. We find that across different contexts, the collected data, which depends on the study objectives, is biased towards certain groups of related subcontexts. For instance, in the tissue context, a large number of subcontexts belong to tissues of the nervous system, while in the cell type context, the majority of subcontexts are related to the immune system. This bias can especially be seen in the cell line context, where nearly all cell lines are derived from cancer cells. Finally, we investigated the correlation between the number of samples or datasets used to generate the co-expression networks and the size of the networks as a potential source of bias. We found no such dependency

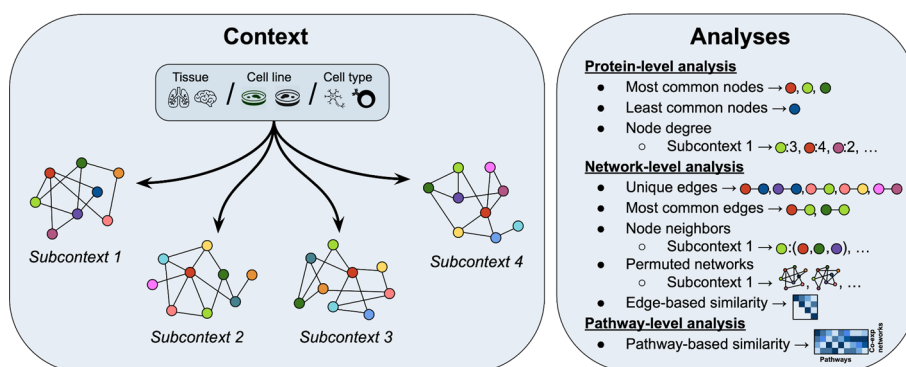


Fig. 2 Overview of analyses conducted across all subcontexts in three different contexts (i.e., tissues, cell lines, and cell types). At the protein-level, patterns surrounding each single node are investigated (“Analyses at the protein-level” section). The network-level analysis focuses on the relations between nodes (or node pairs) (“Analyses at the network-level” section) and the pathway-level analysis leverages defined node and edge sets to gain insights on context-specific co-expression networks (“Mapping co-expression networks to pathway knowledge” section)

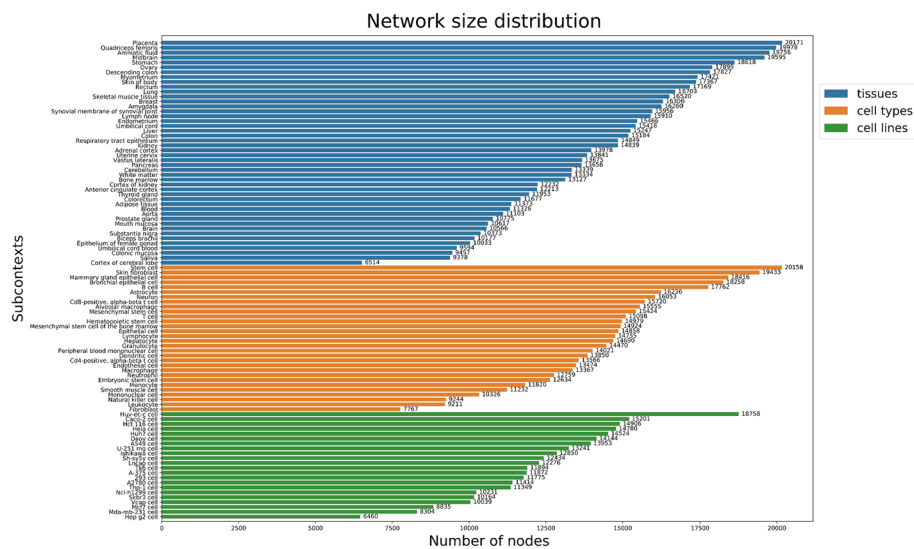


Fig. 3 Distribution of network size for each of the three contexts. Distributions of network size are given as the number of nodes in each subcontext. In the tissue context, the cortex of cerebral lobe network had the fewest number of nodes (i.e., 6514), while the placenta network had the largest number of nodes (i.e., 20,171) across not only all networks of the tissue context, but also across all other contexts. In the cell type context, the fibroblast network had the least number of nodes (i.e., 7767), while the stem cell network had the highest number of nodes (i.e., 20,158). In the cell line context, the HepG2 cell line network had the least number of nodes (i.e., 6460), while the Huv-ec-c cell line network had the largest number of nodes (i.e., 18,758). Generally, the networks within each context tended to vary greatly in size. For example, the tissue context includes networks ranging in size from 6514 to 20,171 nodes

between the number of samples or datasets and the network size (Additional file 2: Fig. S1).

The human interactome we employed (see Methods “Protein—protein interaction network” section), generated in our earlier work [9], contains 8601 nodes and 199,535 edges. These numbers place our interactome on the same scale as other, recently published human interactomes [23], Vinayagam et al. [42]. Nonetheless the size of the interactome, with regard to the number of nodes (proteins), is less than half of the largest co-expression network. This was to be expected, as the majority of proteins measured in transcriptomic experiments have not yet been investigated in the literature and little is known of their functionality. Nodes of the interactome can be visualized in the web application (see “ContNeXt—a web application to explore gene expression patterns across contexts” section) along with their neighbors, betweenness centrality, degree centrality, controllability classification, and information on whether the node is a house-keeping gene.

In order to discern unique features of context-specific co-expression networks which could be of biological significance, we first sought to identify genes known to arise from generic processes whose patterns are more likely to be stable and unaffected by any given context or condition. In particular, we investigated the presence of these, so called, housekeeping genes in each of the co-expression networks, noting that these genes are indicative of shared biology given their role in cell maintenance, and therefore, exhibit constant expression levels across all cells and conditions (Eisenberg and Levanon [7]). Thus, by better understanding which genes have critical roles in basic cell maintenance,

we could better direct our focus in determining genes of interest. The housekeeping genes dataset made available from Eisenberg and Levanon [7] consisted of 3804 genes (Additional file 1: Table S4), 1723 of which were present in the interactome (20% of the overall interactome).

To analyze the structural properties of the interactome, we employed an algorithm (see Method) that has been applied to identify the importance of nodes and edges in biological networks (Additional file 2: Text S1). The results of the controllability analysis indicate that the interactome has 1233 driver nodes with which the network can be controlled. Overall, 74.6% of the nodes were classified as neutral, 16.17% dispensable, and 9.2% indispensable. A list of the full classifications can be seen in Additional file 1: Table S5, with the indispensable nodes listed in Additional file 1: Table S6, and a summary of these nodes can be seen in Table 1. We observed that the indispensable nodes were highly connected, as expected, had the highest average betweenness centrality, and a significant portion (i.e., ~25%) were housekeeping genes. By comparison, neutral nodes were found to have half as many connections and an average betweenness centrality 10 times lower than indispensable nodes. However, the proportion of neutral nodes that were housekeeping genes were comparable to that of the indispensable nodes. By contrast, differences between the dispensable and indispensable nodes were far more pronounced; the average degree of dispensable nodes was only ~6, compared to ~107 for indispensable nodes, while the average betweenness centrality was more than 1000 times lower. Additionally, only ~8% of dispensable nodes were housekeeping genes, compared to roughly a quarter for both indispensable and neutral nodes.

Analyses at the protein-level

We begin by exploring general trends for all co-expression networks of each context at the protein-level by focusing on the most and least common proteins (i.e., present in all or exactly one network within a context). We first used the results of the previously-mentioned controllability analysis of the interactome as well as housekeeping

Table 1 Regarding the interactome controllability, 6417 of the total nodes (74.6%) were classified as neutral; i.e., removing them will have no effect on the number of driver nodes in the network, representing the largest proportion of nodes in the interactome. 1391 (16.17% of the interactome) nodes were dispensable, meaning their removal would decrease the number of driver nodes in the network. Lastly, 793 nodes (9.2% of the interactome) were determined to be indispensable, which caused an increase in the need for driver nodes at their removal. In all three categories (i.e., betweenness centrality, degree, and housekeeping gene proportion), indispensable nodes had the highest value, followed by neutral, and dispensable with the lowest values

	Total number	Scaled betweenness centrality mean	Scaled betweenness centrality median	Scaled betweenness centrality mode	Degree mean	Degree median	Degree mode	Proportion housekeeping gene (%)
Indispensable	793	0.024519	0.006825	0.002642	107.08	60	29	24.59
Dispensable	1391	0.000019	0.000000	0.000000	6.44	4	1	7.84
Neutral	6417	0.004090	0.001101	0.000000	47.56	31	13	22.11

The indispensable nodes are listed in Additional file 1: Table S6. Betweenness centrality scores were scaled between 0 and 1 to facilitate comparability

genes and overlapped them with the most and least common proteins in each context, shown in Additional file 1: Table S7. As summarized in Table 2, of the most common nodes (i.e., proteins that could be found in each network within a given context), we found that the cell type context had the largest number of proteins across all networks (301 proteins), while the tissue network had the fewest (22 proteins). Among the most common nodes, the ratio of housekeeping genes was greater than the proportion of housekeeping genes present in the interactome (i.e., 20%), comprising nearly 50% of the most common nodes in each of the contexts.

Overlap of co-expression networks with the interactome

While only considering the proteins present in the interactome as well as at least one co-expression network, we conducted an in-depth investigation of whether proteins in the co-expression networks of a given context could consistently be identified in the human interactome network. We first noted trends at the protein-level by comparing the most and least common proteins across co-expression networks within a context against the most and least connected proteins of the interactome. As the co-expression network and interactome sizes vastly differed, we studied this overlap considering the top or bottom most proteins in proportions roughly equivalent in size. We selected various cut-offs for each context, corresponding to the number of co-expression networks (see Additional file 2: Text S2 for details on the cut-offs for each context). This ensured the inclusion of either the maximal or minimal possible overlap of the common proteins of the co-expression networks and connected proteins of the interactome, depending on whether our investigation focused on the most commonly or most uniquely occurring proteins, respectively. A detailed list of the resulting overlaps can be seen in Additional file 1: Table S8.

Table 2 Most and least common proteins per context. The most and least common proteins of the co-expression networks (i.e., in all or exactly one network within a context) were overlapped with proteins given distinct classifications from the controllability analysis of the interactome as well as with housekeeping genes. 22 proteins were identified as the most common proteins, that is, found in all 46 co-expression networks of the tissue context. Of the 30 co-expression networks of the cell type context, 301 proteins were found in all of them, while among 22 co-expression networks in the cell line context, 185 proteins were identified in each network. By comparison, no proteins were found to be unique to a single co-expression network in the tissue context, while only one was found in the cell type context belonging to the stem cell co-expression network. On the other hand, 106 least common proteins were found in the cell line context, only one of which is a housekeeping gene and none of which are indispensable

	Tissue context		Cell type context		Cell line context	
Proteins in all co-expression networks	22	2 indispensable 11 housekeeping	301	21 indispensable 180 housekeeping	185	15 indispensable 81 housekeeping
Proteins unique to one co-expression network	0	0 indispensable 0 housekeeping	1	0 indispensable 0 housekeeping	106	0 indispensable 1 housekeeping

A full list of the proteins found in all or in a single network per context can be seen in Additional file 1: Table S7

Most common proteins

First, we focus on the most common proteins. Among the most commonly occurring proteins in the tissue context that overlapped with proteins from the interactome, a number of proteins belonged to the MAPK protein family (Additional file 1: Table S8). Proteins in this family are instrumental in transduction of extracellular signals to cellular responses and complex cellular processes such as apoptosis, development, differentiation, proliferation, and transformation [47]. While only the larger two comparisons in the tissue context (Additional file 2: Fig. S2; lower two diagrams) resulted in an overlap, a significant portion of these overlapping proteins were also indispensable, or housekeeping. Within the large overlaps between the common cell type proteins and most connected interactome proteins (Additional file 2: Fig. S3), a larger proportion of housekeeping genes was found than in any of the contexts studied, with more than half of each overlap being a housekeeping gene (i.e., 50–67%), and more of the proteins are also indispensable.

In cell lines, we observed a substantial overlap of most common proteins that are also found in the interactome overall, including when using the strictest cut-offs, however, significantly less were found to be indispensable or housekeeping than in the tissue and cell type contexts (Additional file 2: Fig. S4). We select a proportional set from each context (400 of the most common proteins per context) to compare their overlaps with the interactome (Additional file 1: Table S9A). The overlaps all had a similar number of proteins in them, between 30 and 37 proteins. Across contexts, there was a similar proportion of the overlap which are indispensable nodes of the interactome (~32% in tissues, 40% in cell types, and ~43% in cell lines). On the other hand, the proportion of housekeeping genes varied more, with 43% of the proteins from the cell line overlap, while tissues and cell types both had more than 60%. Overall, housekeeping genes seem to be best represented in the co-expression networks. We observed a number of proteins in all of the context's overlaps belonging to the Ribosomal protein (RP) family (Additional file 1: Table S9A), from both small and large subunits. RPs are essential in protein synthesis [45]. The tissue overlap had one from large and one from small subunit, the cell type overlap had four from large and one from small subunit, and the cell line overlap had one small subunit RP. We also found that the average number of relations for the proteins in the interactome that overlapped with the approximately top 400 most common proteins in the tissue and cell line networks (~73 and ~72 relations, respectively), was much higher than the average number of relations overall in the interactome (~46 relations). This suggests that the common tissue- and cell line-wide proteins across the co-expression networks are better represented in the scientific literature. In the cell type networks, this average was less high, ~60 relations, but still more than overall in the interactome.

Least common proteins

Next, we investigated the least common proteins in the co-expression networks and their overlap with the least connected proteins in the interactome. This time, the tissue context presented a more consistent overlap while increasing the protein pool, but still a minimal overlap (Additional file 2: Fig. S5). The overlap with the interactome and the cell

type context was about the same as in the tissue context (Additional file 2: Fig. S6). In the cell line context, we found a small, steadily increasing overlap with each interval comparison, which was not the case in the most common proteins (Additional file 2: Fig. S7). The overlap with the interactome in the larger comparisons was roughly the same as in every other context. The minimal overlaps suggest that little is currently known of these proteins. Additionally, we also selected proportional sets of the 400 least common proteins in each context, also occurring the interactome overall against the 400 least connected nodes of the interactome (Additional file 1: Table S9B). The sizes of the overlap didn't vary as much as in the most common and connected comparison, with each context having around 30 proteins in the overlap. As expected, with these overlaps, either one or no proteins are also indispensable or housekeeping. We observe an overwhelming number of proteins belonging to the ZNF protein family in each of the overlaps (i.e., 10/34 (29%) in tissues, 11/33 (33%) in cell types, and 4/27 (15%) in cell lines) (Additional file 1: Table S9B). While ZNFs are widely found in the organism, they play critical roles in specific tissues, and in the development of many diseases [2].

Analyses at the network-level

We first focused on analyzing edges of the co-expression networks, including the unique and most commonly occurring edges within contexts. Additionally, we leveraged prior knowledge from a referential human interactome and studied the correspondence of edges from this network against the strongest pairwise correlations of the co-expression networks. Subsequently, we validated these findings by conducting an equivalent comparison against randomly generated versions of the co-expression networks. Finally, we conducted a similarity analysis on the network edges within each context.

Unique and most commonly occurring edges

We first assessed whether there were any edges specific to particular tissue networks, identifying 45,963,343 unique edges in total (i.e., 49% of all edges). We also identified 34,584,720 unique edges in the cell type context (i.e., 57% of all edges) and 31,941,789 unique edges in the cell line context (i.e., 71% of all edges). These proportions are similar to findings by Stacey et al. [39] who found that over half of edges in several PPI databases are context-specific. Figure 4 illustrates the frequency of unique and common edges in all networks within a context. We find that edges which are common to at least 25% of networks within a context are rare (i.e., between 0.07 and 0.16%), while those which are in at least 75% of networks are nearly negligible (i.e., 33 edges in total for tissues, 9 for cell types, and 4 for cell lines). As only the 1% strongest correlations were selected for each network, it was foreseen that a large number of edges in our resulting co-expression networks would be specific to a single subcontext. Although these unique edges are interesting to explore for a given subcontext (green portions in Fig. 4), given the sheer volume of unique edges, their investigation was outside of the scope of this work.

We hypothesize that these common edges correspond to basal correlations that are not specific as they appear in the majority of networks within one or more contexts. Thus, we analyze the most frequently occurring edges in each of the three contexts. Unsurprisingly, the two housekeeping genes of the tubulin alpha families (i.e., TUBA1C and TUBA1B) are nearly always found to be connected to each other (in 83 out of 98

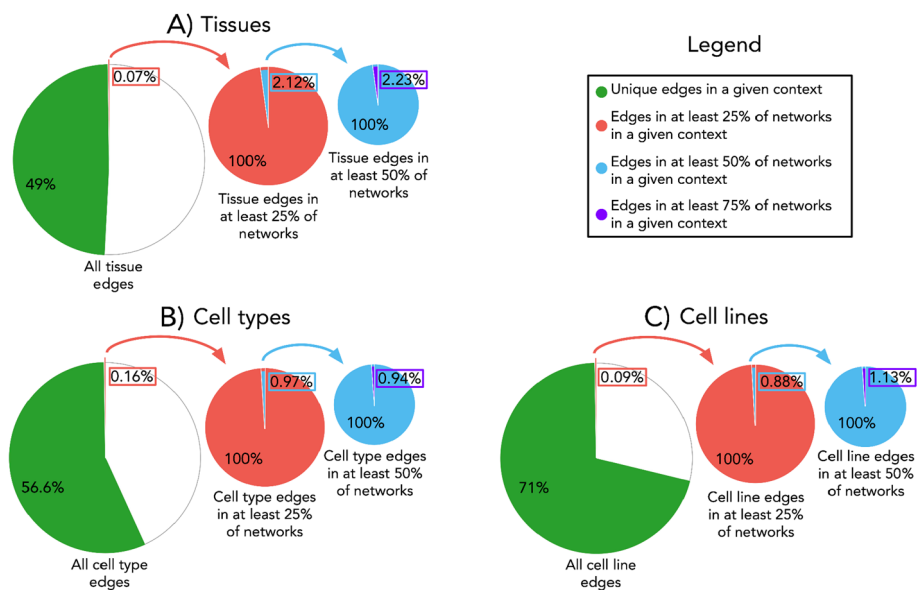


Fig. 4 Frequency of edge occurrence across networks within a context. Proportions of edges are given as those that are unique, or common to varying degrees, in networks within the **A** tissue, **B** cell type, and **C** cell line context. From the total set of edges that occur across all networks within each context, the fraction of edges that are unique (i.e., appear in at most one network within a given context) are shown in green. From this total set of edges, the fraction of those which appear in at least 25% of networks within a given context are magnified in a consecutively smaller pie chart (i.e., predominantly in red). Similarly, those which appear in at least 50% of networks within a given context are magnified and illustrated in a pie chart predominantly in blue. Finally, of this latter group of edges, the fraction of edges that are most common (i.e., appear in at least 75% of all networks within a given context) are highlighted in purple

networks), regardless of context. Additionally, IFITM2 and IFITM3, proteins of the interferon-induced transmembrane family, which play a key role in immune system functions, are also often seen connected to each other in 84 out of 98 networks. Members of the human leukocyte antigens (HLA) protein family are also often interconnected across the cell type and cell line contexts. This is in line with Crow et al. [3] who found that certain gene modules are predictably found across biological conditions, such as those of the immune response. In our previous paper [9], we found that of the most common edges among 63 major diseases, members of the Metallothionein (MT) family of proteins, were in nearly half of these edges. Similarly, here again we observed that a large number of MT proteins share neighbors across networks in every context.

Of the most common edges throughout all contexts (see Additional file 2: Text S3), none were indispensable within the interactome. When widening our search to the top 100,000, we found only seven, three, and one edge in the tissue, cell type, and cell line contexts to be indispensable in the interactome, respectively. Next, these most common edges found in the majority of networks of a given context were compared to the interactome network to identify concordance between the two. We performed a range of comparisons on the most common edges by focusing only on the top 1000 to 10,000 edges, in increments of 1000. Then, the most common edges in each co-expression network were compared to the interactome. Overall, we found little overlap in the most common edges. In the tissue context, we found an overlap of only 5% in the top 1000 most common edges against the interactome, with this overlap decreasing to 4% when considering

the top 10,000 most common edges. In comparison these proportions ranged from ~7 to 3% in the cell type context between the top 1000 and 10,000 most common edges, and 4% to 2% in the cell line context.

The strongest correlations tend to correspond with protein–protein interactions more than expected by chance

In this section, we investigate whether the strongest correlations present in the co-expression networks correspond to PPIs more often than what would be expected by chance. For this purpose, we permuted each co-expression network for each context 1000 times while maintaining the original graph structure (see Methods). We next compared the overlap of edges between these permuted co-expression networks with the human interactome (the results of the first 100 permutations can be seen in Additional file 1: Table S10). Our results show that, on average, the original co-expression networks have 1.55 times as many edges in common with the human interactome as compared to the permuted networks, which exhibited a comparatively low variability in their overlap within a subcontext. Across all contexts, the maximum difference in overlap was for the ovary subcontext, where the original ovary co-expression network had 3.3 times as many edges in common with the interactome as compared to the permuted versions. In comparison, the saliva co-expression network showed the smallest difference in edge overlap between the original and permuted co-expression networks, with the overlap of the interactome with the original co-expression network having only 1.01 times as many edges as the permuted versions on average. Thus, we find that co-expression patterns correspond with PPIs more than expected by chance.

Edge-based similarity across co-expression networks

Next, we investigated edge similarity across networks within a given context. By comparing the co-expression networks to each other rather than just the interactome, we could identify the networks that were most similar edgewise. In the tissue context, two pairs of networks displayed the highest degree of similarity, namely the brain and the cortex of the cerebral lobe, and the colon and the rectum (Fig. 5A). This finding was not surprising given that these pairs of tissues are anatomically related (i.e., both are of the brain or the colorectum). The cell line context had a few standout pairs of networks which had the highest degree of similarity (Fig. 5B). Specifically, the highest similarity was between two different human breast cancer cell lines: MDA-MB-231 and MCF7. Additionally, the MCF7 cell line again had a high similarity with a human colon cancer cell line, HCT 116. On the other hand, in the cell type context, rather than specific pairs showing the highest similarity with each other, a few selected subcontexts had a high similarity with most of the other networks overall (Fig. 5C). In particular, the peripheral blood mononuclear cell network showed high similarity with its more specific cell type networks, including monocytes, T cells, and lymphocytes. Overall, these results lend support to how network similarity can reflect similarity across related cell types, tissues, or cell lines.

Mapping co-expression networks to pathway knowledge

Lastly, we attempted to establish patterns across co-expression networks at a pathway-level by overlaying pathway knowledge with the co-expression networks. If a given

Pairwise network similarity across each context based on edge overlap

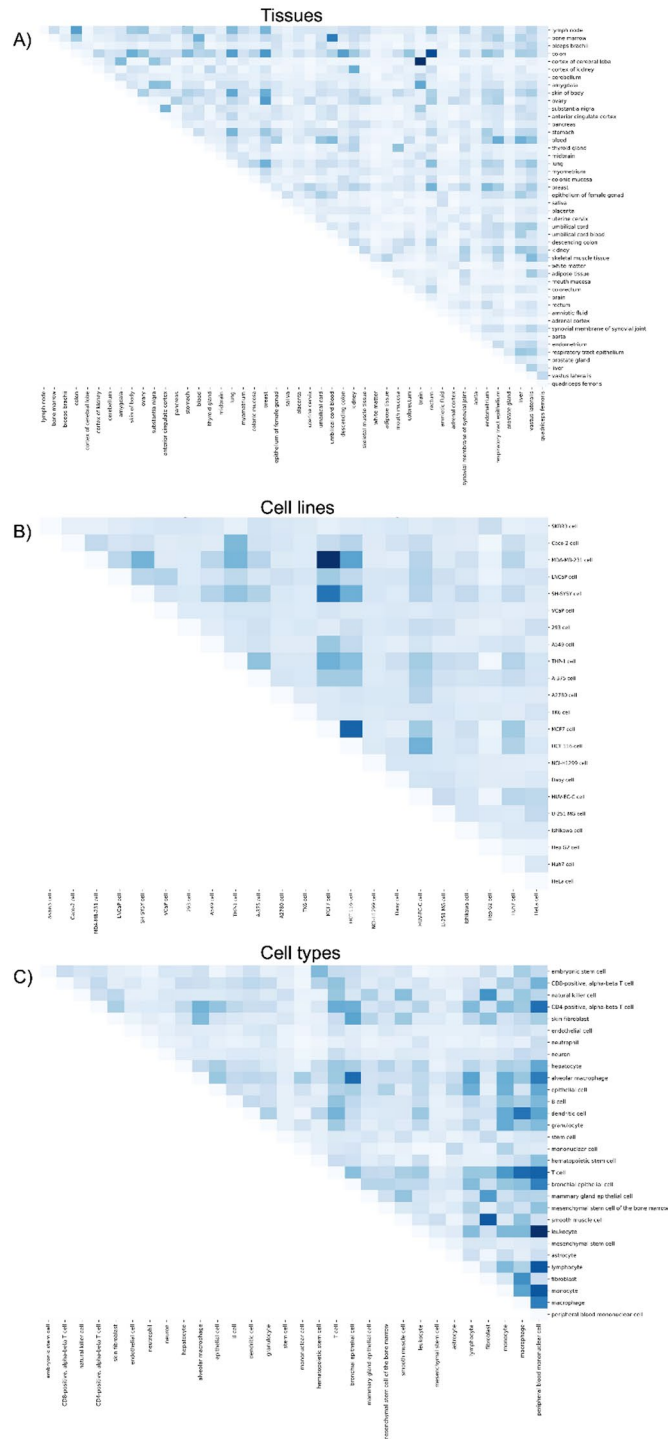


Fig. 5 Pairwise co-expression network similarity across contexts. For each pair of co-expression networks within a given context, edge overlap was calculated as a measure of similarity between networks for the **A** tissue, **B** cell line, and **C** cell type contexts. A high quality version of the figure is available at <https://github.com/ContNexT/scripts/blob/main/figures/figure5.pdf>

pathway is related to a specific network (e.g., fatty acid metabolism pathway and the liver co-expression network), we would expect that the proteins in the pathway would be strongly correlated in the co-expression network. Furthermore, we assume that, given a set of highly co-expressed genes of which a majority are involved in a particular pathway, the remaining genes may be functionally relevant to the pathway as well. We therefore seek to identify the pathways associated with networks from each of the investigated contexts. Using the KEGG database [15], we mapped pathway knowledge to co-expression networks according to Eq. 1 (see Methods).

We found several groups of tissues that had high similarities with pathways related to the given tissues (Fig. 6). For instance, the two tissue networks corresponding to cortex of cerebral lobe and brain shared a large group of pathways exhibiting a high degree of similarity, including nine synaptic pathways (Fig. 6; green oval) (Additional file 1: Table S11). Furthermore, the three networks for liver, cortex of kidney, and kidney also had the highest level of similarity with numerous pathways, including eight involving the regulation of fatty acids as well as 11 involving amino acid metabolism and degradation (Fig. 6; red oval) (Additional file 1: Table S12). Not surprisingly, the adipose tissue network also showed the highest similarity with adipose-related pathways, such as adipocytokine signaling pathway and regulation of lipolysis in adipocytes pathway.

In the cell type context, while no groups of network shared distinct pathways among them, we found three cell types having distinct groups of pathways with very high similarity unique to a single network. For example, a number of pathways showed a high degree of similarity to the neutrophil co-expression network (Additional file 2: Fig. S8; red oval), namely, 11 that regulate the immune response (Additional file 1: Table S13). Additionally, the co-expression network for hepatocytes, the primary cell type of the liver, had the highest level of similarity with many pathways (Additional file 2: Fig. S8; yellow oval), including six involving basic liver function as well as many metabolic pathways, particularly 10 pertaining to amino acids metabolism and seven for other specific molecules (Additional file 1: Table S14). Lastly, we found an additional group of pathways that were exclusively similar to one network, namely the neuron (Additional file 2: Fig. S8; green oval). Specifically, this included five pathways related to neurotransmitter systems, long-term depression, and pathways related to addiction (Additional file 1: Table S15).

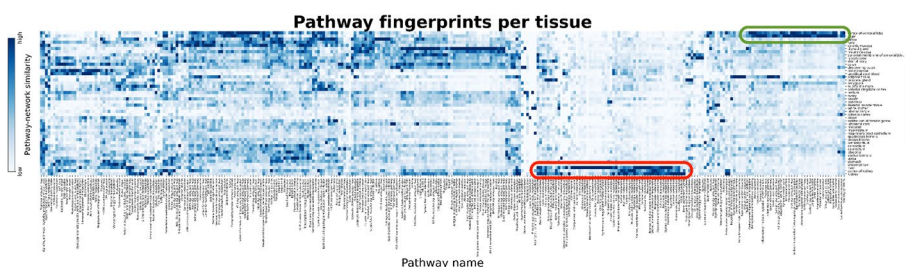


Fig. 6 Similarity between tissue-specific co-expression networks and KEGG pathways. The similarity between a particular pathway and a co-expression network is defined as the percentage of pairwise combinations of proteins of a given KEGG pathway that can be found in a co-expression network as edges. Light blue corresponds to a lower similarity, while dark blue corresponds to a high similarity. A high quality version of this figure is available at https://github.com/ContNeXt/scripts/blob/main/figures/figure6_highquality.pdf and can also be visualized in the web application

Analogous to the cell type context, while related groups of networks from the cell line context were not found to be similar to related groups of pathways (Additional file 2: Fig. S9), several individual cell lines were observed to be highly similar to a group of pathways. However, these pathways were not necessarily unique to the cell line, showing some similarity with other cell lines as well. Interestingly, we found a large group of pathways (i.e., 70 in total) with consistently high similarity with nearly all cell lines, with the exception of the THP-1 cell line (Additional file 2: Fig. S9; green rectangle). These include 24 different signaling pathways and 16 different cancer pathways (Additional file 1: Table S16). Notably, we found a group of pathways that were distinctly similar to two cell lines (i.e., A549 and TK6). Specifically, 14 pathways showed a high degree of similarity to the A549 cell line co-expression network (Additional file 2: Fig. S9; yellow oval). This cell line originated from adenocarcinomic human alveolar basal epithelial cells from lung cancer and is used as a model for drug metabolism [10]. Of these 14 pathways that, on average, showed the highest similarity to this cell line relative to the others, eight were pathways involving metabolism and three were pathways related to compound biosynthesis (Additional file 1: Table S17). Similarly, we identified a group of pathways which showed a higher similarity to the TK6 cell line, originating from a human B lymphoblastoid cell [36], over all other cell lines (Additional file 2: Fig. S9; red oval), including five signaling pathways (Additional file 1: Table S18).

ContNeXt—a web application to explore gene expression patterns across contexts

To provide access to the co-expression networks and analyses presented in this work, we have developed ContNeXt, a web application that facilitates the large-scale exploration and analysis of transcriptomic patterns across multiple contexts. The main page of the web application allows users to search co-expression patterns for a given node in a particular context or browse and query specific nodes in a certain subcontext (Fig. 7A). With interactive network visualizations, users can explore these patterns and employ functionalities such as filtering or search boxes (Fig. 7B). Similarly, the heatmaps presented in this work can be interactively explored through the web application (Fig. 7C). Finally, both the processed data and networks can be downloaded directly from the web application.

Discussion

We have presented a large-scale network-based approach that aims at revealing common and specific biological processes and mechanisms across contexts by identifying transcriptional patterns that are unique to various cell types, tissues, and cell lines, as well as patterns which are consistent across them. In order to do so, we constructed co-expression networks to capture the strongest correlations observed in 98 specific subcontexts belonging to these three biological contexts (i.e., tissues, cell types, and cell lines) and conducted a series of analyses at the protein, network, and pathway levels. Finally, we developed a web application to enable users to query and display these networks and ultimately, explore shared and distinct co-expression patterns for multiple contexts.

We believe that one strength of our work is its robustness, as we have systematically leveraged hundreds of curated datasets, thereby ensuring a diverse sample of experiments conducted in similar settings whilst applying a common preprocessing and

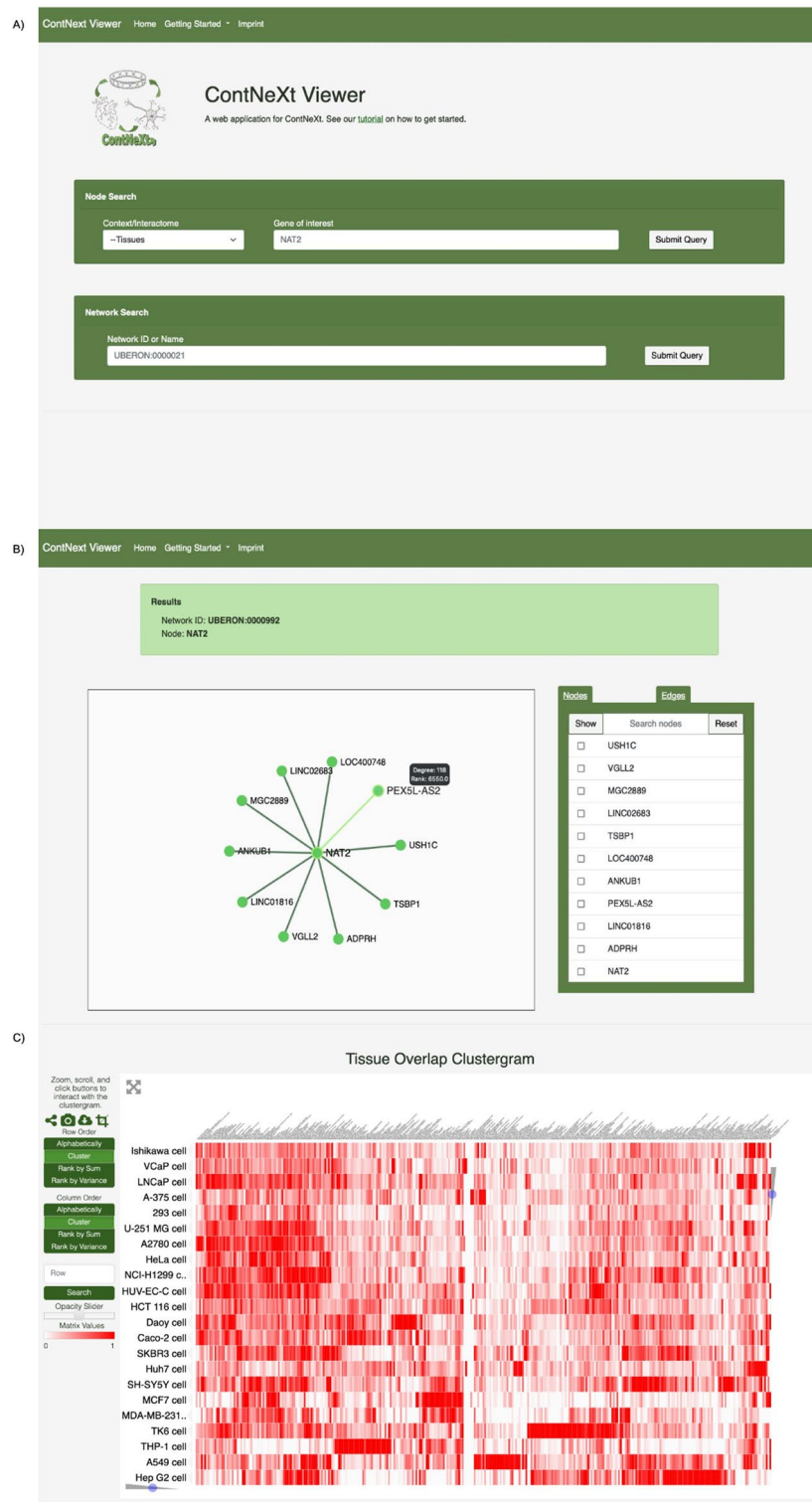


Fig. 7 ContNeXt web application. **A** Main page. Users can query for specific genes or directly explore the networks of a given context. **B** Network page. Users can explore and navigate through the neighbors of a specific gene for each network. **C** Heatmap visualization. Heatmaps presented in this work can be interactively viewed to investigate pairwise co-expression network-based similarity as well as pathway- co-expression network -based similarity

analysis pipeline. However, although we applied a conservative inclusion/exclusion criteria, we cannot assume that every dataset in the same (sub)context is equivalent and thus, some of the patterns identified may be dataset-specific. To account for this factor and reduce noise and variability across datasets, we focused on the 1% strongest correlations, keeping in mind that the choice of cut-off can influence the resulting co-expression network [44], and also constrained our analysis to subcontexts with a large number of samples. Still, independently of this minimum criteria, there are differences in the number of datasets per subcontext that could lead to variability for specific subcontexts with a small sample size. Another limitation is that we have exclusively relied on the platform with a large number of datasets in the Gemma database. Similarly, we also employed Gemma's context annotations to classify the datasets. While it is technically possible to include more platforms in our analysis as well as annotate datasets from other databases, each additional platform would require its own independent processing pipeline and a significant curation effort. Furthermore, in the cell line context, it is important to note that the majority of cell lines originate from widely used immortal cancer cell lines, which might differ from the normal human cells used for the cell type and tissue contexts. Finally, we would like to remark on two other limitations of our analysis. Firstly, while we employed a large and high-quality version of the protein–protein human interactome, some parts of the graphs are more dense than others as some proteins are under-studied [35]. Secondly, some of the analyses are influenced by the size of the co-expression networks (Fig. 3), as the fewer nodes a network has, the more dense it is due to the larger amount of connections between its nodes.

Lastly, we would like to mention some of the prospects we foresee for future work. Firstly, by further incorporating single-cell experiment datasets, we can potentially identify more granular patterns. Additional single-cell RNA-seq datasets can be included in our work to verify whether the observed tissue-specific transcriptional patterns are indeed characteristic to specific tissues, or are influenced by their cellular composition, as observed by Farahbod and Pavlidis [8]. While this large-scale exercise is not feasible at the moment due to the lack of available data of this kind, we expect that it could be conducted in future. Secondly, disease-specific gene expression datasets can be exploited to compare disease-specific signatures with the ones observed in a related normal tissue or cell type in order to identify the biological processes and pathways that are dysregulated in the disease context. Thirdly, as demonstrated by Azevedo et al. [1] and Sealfon et al. [37], machine learning models could be trained on the generated co-expression networks to classify signatures coming from new samples into a particular context given its specific characteristics.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12859-022-04765-0>.

Additional file 1 Supplementary Tables.

Additional file 2 Supplementary File including supplementary figure and tables.

Acknowledgements

We would like to thank the entire Gemma team, especially Paul Pavlidis, for their support using their tool. Furthermore, we would like to thank André Gemünd for his technical assistance.

Author contributions

DDF and SM conceived and designed the study. RQF and TR processed the transcriptomic datasets. RQF implemented the methodology and analyzed the results supervised by SM and DDF. SDS implemented the web application. ATK, MHA and DDF acquired the funding. RQF, SM, and DDF wrote the manuscript. ATK and MHA reviewed the manuscript. All authors read and approved final manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL. This work was developed in the Fraunhofer Cluster of Excellence "Cognitive Internet Technologies". This work is supported by the German Federal Ministry of Education and Research (BMBF, grant 01ZX1904C).

Availability of data and materials

All data supporting the conclusions of this article are available at <https://zenodo.org/record/5831786> and scripts can be found at <https://github.com/ContNeXt/scripts>. ContNeXt and its source code are available at <https://contnext.scai.fraunhofer.de> and https://github.com/ContNeXt/web_app, respectively.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

DDF received salary from Enveda Biosciences. Other authors do not declare any competing interests.

Received: 11 February 2022 Accepted: 3 June 2022

Published online: 15 June 2022

References

1. Azevedo T, Dimitri GM, Lió P, Gamazon ER. Multilayer modelling of the human transcriptome and biological mechanisms of complex diseases and traits. *NPJ Sys Biol Appl*. 2021;7(1):1–13. <https://doi.org/10.1038/s41540-021-00186-6>.
2. Cassandri M, Smirnov A, Novelli F, Pitolli C, Agostini M, Malewicz M, et al. Zinc-finger proteins in health and disease. *Cell Death Discov*. 2017;3(1):1–12. <https://doi.org/10.1038/cddiscovery.2017.71>.
3. Crow M, Lim N, Ballouz S, Pavlidis P, Gillis J. Predictability of human differential gene expression. *Proc Natl Acad Sci*. 2019;116(13):6491–500. <https://doi.org/10.1073/pnas.1802973116>.
4. Diehl AD, Meehan TF, Bradford YM, Brush MH, Dahdul WM, Dougall DS, et al. The cell ontology 2016: enhanced content, modularization, and ontology interoperability. *J Biomed Semant*. 2016;7(1):1–10. <https://doi.org/10.1186/s13326-016-0088-7>.
5. Dobrin R, Zhu J, Molony C, Argman C, Parrish ML, Carlson S, Allan MF, Pomp D, Schadt EE. Multi-tissue coexpression networks reveal unexpected subnetworks associated with disease. *Genome Biol*. 2009;10(5):1–3. <https://doi.org/10.1186/gb-2009-10-5-r55>.
6. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*. 2002;30(1):207–10. <https://doi.org/10.1093/nar/30.1.207>.
7. Eisenberg E, Levanon EY. Human housekeeping genes, revisited. *Trends Genet*. 2013;29(10):569–74. <https://doi.org/10.1016/j.tig.2013.05.010>.
8. Farahbod M, Pavlidis P. Untangling the effects of cellular composition on coexpression analysis. *Genome Res*. 2020;30(6):849–59. <https://doi.org/10.1101/gr.256735.119>.
9. Figueiredo RQ, Raschka T, Kodamullil AT, Hofmann-Apitius M, Mubeen S, Domingo-Fernández D. Towards a global investigation of transcriptomic signatures through co-expression networks and pathway knowledge for the identification of disease mechanisms. *Nucleic Acids Res*. 2021;49(14):7939–53. <https://doi.org/10.1093/nar/gkab556>.
10. Foster KA, Oster CG, Mayer MM, Avery ML, Audus KL. Characterization of the A549 cell line as a type II pulmonary epithelial cell model for drug metabolism. *Exp Cell Res*. 1998;243(2):359–66. <https://doi.org/10.1006/excr.1998.4172>.
11. Hagberg AA, Schult DA, Swart PJ. Exploring network structure, dynamics, and function using NetworkX. In: *Proceedings of the 7th Python in Science Conference (SciPy2008)*; 2008. Pp. 11–5.
12. Hanhijärvi S, Garriga, GC, Puolamäki K. Randomization techniques for graphs. In: *Proceedings of the 2009 SIAM International Conference on Data Mining*; 2009. pp. 780–91. <https://doi.org/10.1137/1.9781611972795.67>
13. Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, D'Eustachio P. The reactome pathway knowledgebase. *Nucleic Acids Res*. 2020;48(D1):D498–503. <https://doi.org/10.1093/nar/gkz1031>.
14. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007;8(1):118–27. <https://doi.org/10.1093/biostatistics/kxj037>.
15. Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M. KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res*. 2021;49(D1):D545–51. <https://doi.org/10.1093/nar/gkaa970>.
16. Kitsak M, Sharma A, Menche J, Guney E, Ghiassian SD, Loscalzo J, Barabási AL. Tissue specificity of human disease module. *Sci Rep*. 2016;6(1):1–12. <https://doi.org/10.1038/srep35241>.
17. Koussounadis A, Langdon SP, Um IH, Harrison DJ, Smith VA. Relationship between differentially expressed mRNA and mRNA-protein correlations in a xenograft model system. *Sci Rep*. 2015;5(1):1–9. <https://doi.org/10.1038/srep10775>.
18. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinform*. 2008;9(1):1–13. <https://doi.org/10.1186/1471-2105-9-559>.

19. Lee YF, Lee CY, Lai LC, Tsai MH, Lu TP, Chuang EY. Cell Express: a comprehensive microarray-based cancer cell line and clinical sample gene expression analysis online system. Database. 2018. <https://doi.org/10.1093/database/bax101>.
20. Lee J, Shah M, Ballouz S, Crow M, Gillis J. CoCoNet: conserved and comparative co-expression across a diverse set of species. *Nucleic Acids Res.* 2020;48(W1):W566–71. <https://doi.org/10.1093/nar/gkaa348>.
21. Lim N, Tesar S, Belmadani M, Poirier-Morency G, Mancarci BO, Sicherman J, et al. Curation of over 10,000 transcriptomic studies to enable data reuse. Database. 2021. <https://doi.org/10.1093/database/baab006>.
22. Liu YY, Slotine JJ, Barabási AL. Controllability of complex networks. *Nature.* 2011;473(7346):167–73. <https://doi.org/10.1038/nature10011>.
23. Luck K, Kim DK, Lambourne L, Spirohn K, Begg BE, Bian W, et al. A reference map of the human binary protein interactome. *Nature.* 2020;580(7803):402–8. <https://doi.org/10.1038/s41586-020-2188-x>.
24. McKenzie AT, Wang M, Hauberg ME, Fullard JF, Kozlenkov A, Keenan A, et al. Brain cell type specific gene expression and co-expression network architectures. *Sci Rep.* 2018;8(1):1–9. <https://doi.org/10.1038/s41598-018-27293-5>.
25. Mungall CJ, Torniai C, Gkoutos GV, Lewis SE, Haendel MA. Uberon, an integrative multi-species anatomy ontology. *Genome Biol.* 2012;13(1):1–20. <https://doi.org/10.1186/gb-2012-13-1-r5>.
26. Nusinow DP, Szpyt J, Ghandi M, Rose CM, McDonald ER III, Kalocsay M, et al. Quantitative proteomics of the cancer cell line encyclopedia. *Cell.* 2020;180(2):387–402. <https://doi.org/10.1016/j.cell.2019.12.023>.
27. Obayashi T, Kagaya Y, Aoki Y, Tadaka S, Kinoshita K. COXPRESdb v7: a gene coexpression database for 11 animal species supported by 23 coexpression platforms for technical evaluation and evolutionary inference. *Nucleic Acids Res.* 2019;47(D1):D55–62. <https://doi.org/10.1093/nar/gky1155>.
28. Oldham MC, Konopka G, Iwamoto K, Langfelder P, Kato T, Horvath S, Geschwind DH. Functional organization of the transcriptome in human brain. *Nat Neurosci.* 2008;11(11):1271–82. <https://doi.org/10.1038/nn.2207>.
29. Perkins AD, Langston MA. Threshold selection in gene co-expression networks using spectral graph theory techniques. *BMC Bioinform.* 2009;10(11):1–11. <https://doi.org/10.1186/1471-2105-10-S11-S4>.
30. Pierson E, GTEx Consortium, Koller D, Battle A, Mostafavi S. Sharing and specificity of co-expression networks across 35 human tissues. *PLoS Comput Biol.* 2015;11(5):e1004220. <https://doi.org/10.1371/journal.pcbi.1004220>.
31. Rachlin J, Cohen DD, Cantor C, Kasif S. Biological context networks: a mosaic view of the interactome. *Mol Syst Biol.* 2006;2(1):66. <https://doi.org/10.1038/msb4100103>.
32. Romero IG, Ruvinsky I, Gilad Y. Comparative studies of gene expression and the evolution of gene regulation. *Nat Rev Genet.* 2012;13(7):505–16. <https://doi.org/10.1038/nrg3229>.
33. Rung J, Brazma A. Reuse of public genome-wide gene expression data. *Nat Rev Genet.* 2013;14(2):89–99. <https://doi.org/10.1038/nrg3394>.
34. Sarntivijai S, Lin Y, Xiang Z, Meehan TF, Diehl AD, Vempati UD, et al. CLO: the cell line ontology. *J Biomed Semant.* 2014;5(1):1–10. <https://doi.org/10.1186/2041-1480-5-37>.
35. Schaefer MH, Serrano L, Andrade-Navarro MA. Correcting for the study bias associated with protein–protein interaction measurements reveals differences between protein degree distributions from different cancer types. *Front Genet.* 2015;6:260. <https://doi.org/10.3389/fgene.2015.00260>.
36. Schwartz JL, Jordan R, Evans HH, Lenarczyk M, Liber HL. Baseline levels of chromosome instability in the human lymphoblastoid cell TK6. *Mutagenesis.* 2004;19(6):477–82. <https://doi.org/10.1093/mutage/geh060>.
37. Sealfon RS, Wong AK, Troyanskaya OG. Machine learning methods to model multicellular complexity and tissue specificity. *Nat Rev Mater.* 2021. <https://doi.org/10.1038/s41578-021-00339-3>.
38. Sonawane AR, et al. Understanding tissue-specific gene regulation. *Cell Rep.* 2017;21(4):1077–88. <https://doi.org/10.1016/j.celrep.2017.10.001>.
39. Stacey RG, Skinnider MA, Chik JHL, Foster LJ. Context-specific interactions in literature-curated protein interaction databases. *BMC Genom.* 2018;19(1):1–10. <https://doi.org/10.1186/s12864-018-5139-2>.
40. Trapotsi MA, Hosseini-Gerami L, Bender A. Computational analyses of mechanism of action (MoA): data, methods and integration. *RSC Chem Biol.* 2022. <https://doi.org/10.1039/D1CB00069A>.
41. The Gene Ontology Consortium. The gene ontology resource: enriching a GOLD mine. *Nucleic Acids Res.* 2021;49(D1):D325–34. <https://doi.org/10.1093/nar/gkaa1113>.
42. Vinayagam A, Gibson TE, Lee HJ, Yilmazel B, Roesel C, Hu Y, et al. Controllability analysis of the directed human protein interaction network identifies disease genes and drug targets. *Proc Natl Acad Sci.* 2016;113(18):4976–81. <https://doi.org/10.1073/pnas.1603992113>.
43. Whitehead A, Crawford DL. Variation in tissue-specific gene expression among natural populations. *Genome Biol.* 2005;6(2):1–14. <https://doi.org/10.1186/gb-2005-6-2-r13>.
44. Yip AM, Horvath S. Gene network interconnectedness and the generalized topological overlap measure. *BMC Bioinform.* 2007;8(1):1–14. <https://doi.org/10.1186/1471-2105-8-22>.
45. Yoshihama M, Uechi T, Asakawa S, Kawasaki K, Kato S, Higa S, Maeda N, Minoshima S, Tanaka T, Shimizu N, Kenmochi N. The human ribosomal protein genes: sequencing and comparative analysis of 73 genes. *Genome Res.* 2002;12(3):379–90. <https://doi.org/10.1101/gr.214202>.
46. Yu K, Chen B, Aran D, Charalel J, Yau C, Wolf DM, et al. Comprehensive transcriptomic analysis of cell lines as models of primary tumors across 22 tumor types. *Nat Commun.* 2019;10(1):1–11. <https://doi.org/10.1038/s41467-019-11415-2>.
47. Zhang W, Liu HT. MAPK signal pathways in the regulation of cell proliferation in mammalian cells. *Cell Res.* 2002;12(1):9–18. <https://doi.org/10.1038/sj.cr.7290105>.
48. Zoubarev A, Hamer KM, Keshav KD, McCarthy EL, Santos JRC, Van Rossum T, et al. Gemma: a resource for the reuse, sharing and meta-analysis of expression profiling data. *Bioinformatics.* 2012;28(17):2272–3. <https://doi.org/10.1093/bioinformatics/bts430>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

A.6 MultiPaths: a Python framework for analyzing multi-layer biological networks using diffusion algorithms

Reprinted with permission from “Marín-Llaó J., Mubeen S., Perera-Lluna A., Hofmann-Apitius M., Picart-Armada S., and Domingo-Fernández D. (2021). MultiPaths: a Python framework for analyzing multi-layer biological networks using diffusion algorithms. *Bioinformatics*, 37(1): 137-139”.

Copyright © Marín-Llaó, J., *et al.*, 2021.

Systems biology

MultiPaths: a Python framework for analyzing multi-layer biological networks using diffusion algorithms

Josep Marín-Llaó^{1,2}, Sarah Mubeen ^{1,3}, Alexandre Perera-Lluna ²,
Martin Hofmann-Apitius ¹, Sergio Picart-Armada ^{2,*†} and
Daniel Domingo-Fernández ^{1,3,*†}

¹Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Sankt Augustin 53757, Germany, ²B2SLab, Departament d'Enginyeria de Sistemes, Automàtica i Informàtica Industrial, Universitat Politècnica de Catalunya, CIBER-BBN, Barcelona 08028, Spain and ³Fraunhofer Center for Machine Learning, Germany

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the last two authors should be regarded as Joint Last Authors.

Associate Editor: Lenore Cowen

Received on August 12, 2020; revised on November 23, 2020; editorial decision on December 11, 2020; accepted on December 14, 2020

Abstract

Summary: High-throughput screening yields vast amounts of biological data which can be highly challenging to interpret. In response, knowledge-driven approaches emerged as possible solutions to analyze large datasets by leveraging prior knowledge of biomolecular interactions represented in the form of biological networks. Nonetheless, given their size and complexity, their manual investigation quickly becomes impractical. Thus, computational approaches, such as diffusion algorithms, are often employed to interpret and contextualize the results of high-throughput experiments. Here, we present MultiPaths, a framework consisting of two independent Python packages for network analysis. While the first package, DiffuPy, comprises numerous commonly used diffusion algorithms applicable to any generic network, the second, DiffuPath, enables the application of these algorithms on multi-layer biological networks. To facilitate its usability, the framework includes a command line interface, reproducible examples and documentation. To demonstrate the framework, we conducted several diffusion experiments on three independent multi-omics datasets over disparate networks generated from pathway databases, thus, highlighting the ability of multi-layer networks to integrate multiple modalities. Finally, the results of these experiments demonstrate how the generation of harmonized networks from disparate databases can improve predictive performance with respect to individual resources.

Availability and implementation: DiffuPy and DiffuPath are publicly available under the Apache License 2.0 at <https://github.com/multipaths>.

Contact: sergi.picart@upc.edu or daniel.domingo.fernandez@scai.fraunhofer.de

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

Emergent properties of biological processes primarily arise from complex interactions linking physical entities which, in turn, can build up complex biological networks, such as metabolic, signaling and regulatory. The use of these networks has become commonplace for a variety of analytic tasks, yet integrated networks have been shown to be more robust resources for analytic usage (Huang *et al.*, 2018). Thus, several frameworks, such as Bio2RDF (Belleau *et al.*, 2008), have been proposed to facilitate the integration of these networks from heterogeneous sources.

Numerous methods for network analysis derived from graph theory have been adapted for a broad range of applications in the biomedical domain including target prioritization, gene prediction and patient stratification (Barabási *et al.*, 2011; Pai *et al.*, 2019). Amongst these methods, network propagation or diffusion, in particular, comprises a broad family of algorithms that infer node labels based on the sharing of labels through network connections (Cowen *et al.*, 2017).

Though a wide variety of algorithms exist, user-friendly software that can enable researchers to implement and compare several methods are lacking. Not only does this impede their adoption and

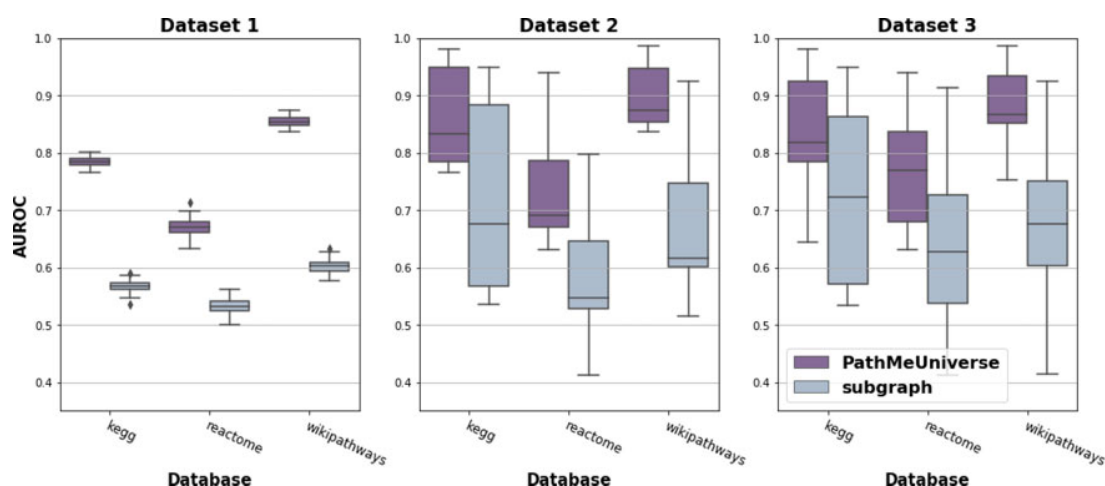


Fig. 1. Prediction performance of raw diffusion over the integrated PathMe network and the correspondent database subgraph for three multi-omics datasets. Each box plot shows the distribution of the area under the ROC curves (AUROCs) over 100 repeated holdout validations. Details on the evaluation can be found in the Supplementary Text

reproducibility but it also compels researchers to re-implement the algorithms for their particular needs. While recently, the R packages *diffuStats* and *RANKS* (Picart-Armada *et al.*, 2018; Valentini *et al.*, 2016) have addressed this issue, a framework that offers a pipeline to build harmonized networks from biological databases along with an array of ready-to-use diffusion algorithms has yet to be established.

Here, we present *MultiPaths*, a Python framework for the analysis of multi-omics data by classical and statistically normalized diffusion algorithms on harmonized networks from custom or predefined selections of biological databases. We demonstrate how *MultiPaths* enables contextualizing multi-omics experiments by presenting an application scenario on multiple datasets containing transcriptomics, metabolomics and miRNomics data.

2 Implementation

The *MultiPaths* framework contains two independent Python packages: *DiffuPy* and *DiffuPath*. While *DiffuPy* is specifically designed for the implementation of diffusion algorithms, *DiffuPath* is capable of both generating harmonized biological networks, and running the algorithms over these networks. Their functionalities can be accessed programmatically and via a command line interface (CLI) for nonbioinformaticians. Their modular design eases the inclusion of network resources and algorithms in future releases.

2.1 DiffuPy

The first of the two packages in the framework, *DiffuPy*, enables propagating user-defined labels, either as lists of entities or lists of entities with their corresponding quantitative values, on a user-defined network. *DiffuPy* comprises four diffusion scores and five graph kernels that can be run on generic networks on different formats (Supplementary Text).

2.2 DiffuPath

The second package, *DiffuPath*, wraps the generic diffusion algorithms from *DiffuPy* and applies them to biological networks. To that end, *DiffuPath* comprises a comprehensive pipeline that extends from the generation of harmonized networks from multiple biological databases to the visualization and analysis of the diffusion results (Supplementary Text). The pipeline provides a user-friendly CLI that enables users to create customized networks from a pool of databases or predefined collections based on their input data, directly run diffusion algorithms on these networks, and analyze them in a few commands. Finally, we would like to note that, while *DiffuPath* already includes a wide range of databases, the framework supports

the integration of any number of databases in standard network formats.

3 Application

To demonstrate the framework, we run various diffusion algorithms from *DiffuPy* on four networks corresponding to four pathway databases generated through *DiffuPath*. The input labels for the diffusion derive from three independent datasets containing differential entities from three -omics modalities: transcriptomics, metabolomics and miRNomics. The four networks consist of three well-established pathway databases: KEGG, Reactome and WikiPathways (Fabregat *et al.*, 2018; Kanehisa *et al.*, 2017; Slenter *et al.*, 2018) as well as their combined representation, PathMe (Domingo-Fernández *et al.*, 2019). Our hypothesis is that by integrating the three resources, PathMe covers a larger scale of interactions and entities as well as a broader range of interaction and modality types which can ultimately serve to improve prediction performance.

For each of the three datasets which investigated specific biological processes, we compared the prediction performance of the various diffusion algorithms in identifying genes, metabolites and miRNAs (for details see Supplementary Text Section S4: Case scenario). This was repeated for each of the four networks and the performance was evaluated using a repeated holdout approach. For the raw diffusion scores, the distribution of area under the ROC curve (AUROC) scores indicated a significant improvement in prediction performance of the integrated multi-layer network over each of the individual databases (Fig. 1).

4 Discussion

This work has presented the first Python framework that implements numerous diffusion algorithms along with a pipeline to build customized harmonized networks from multiple biological databases. The importance of this integration is highlighted by our three case scenarios where a harmonized network leverages three -omics modalities (Di Nanni *et al.*, 2020) to increase predictivity in line with Huang *et al.* (2018). Furthermore, the integrated networks contain additional entities like biological processes and clinical readouts (e.g. symptoms and diseases), allowing a rich contextualization of the experimental readouts (Supplementary Text). This case scenario demonstrates the utility of diffusion algorithms to provide the interpretation of biological networks in the context of pathways, and thereby, elucidate the properties of biological processes underlying these networks. Additionally, users can conduct analyses from

biological networks to generic networks from other fields (e.g. social media) as well as incorporate additional kernels or diffusion algorithms to DiffuPy. As a final remark, although large-scale and integrated multi-layer networks can improve prediction performance, greater computational power is required as the size of a network grows.

Funding

This work was developed in the Fraunhofer Cluster of Excellence ‘Cognitive Internet Technologies’ and the DPI2017-89827-R, Networking Biomedical Research Centre in the subject area of Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN).

Conflict of Interest: none declared.

References

- Barabási,A.L. *et al.* (2011) Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.*, **12**, 56–68.
- Belleau,F. *et al.* (2008) Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *J. Biomed. Inf.*, **41**, 706–716.
- Cowen,L. *et al.* (2017) Network propagation: a universal amplifier of genetic associations. *Nat. Rev. Genet.*, **18**, 551–562.
- Di Nanni,N. *et al.* (2020) Network diffusion promotes the integrative analysis of multiple omics. *Front. Genet.*, **11**, 106.
- Domingo-Fernández,D. *et al.* (2019) PathMe: merging and exploring mechanistic pathway knowledge. *BMC Bioinformatics*, **20**, 243.
- Fabregat,A. *et al.* (2018) The Reactome pathway knowledgebase. *Nucleic Acids Res.*, **46**, D649–D655.
- Huang,J.K. *et al.* (2018) Systematic evaluation of molecular networks for discovery of disease genes. *Cell Syst.*, **6**, 484–495.
- Kanehisa,M. *et al.* (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.*, **45**, D353–D361.
- Pai,S. *et al.* (2019) netDx: interpretable patient classification using integrated patient similarity networks. *Mol. Syst. Biol.*, **15**, e8497.
- Picart-Armada,S. *et al.* (2018) diffuStats: an R package to compute diffusion-based scores on biological networks. *Bioinformatics*, **34**, 533–534.
- Slenter,N. *et al.* (2018) WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res.*, **46**, D661–D667.
- Valentini,G. *et al.* (2016) RANKS: a flexible tool for node label ranking and classification in biological networks. *Bioinformatics*, **32**, 2872–2874.

A.7 Drug2ways: reasoning over causal paths in biological networks for drug discovery

Reprinted with permission from “Rivas-Barragan D., Mubeen S., Guim Bernat F., Hofmann-Apitius M., and Domingo-Fernández D. (2020). Drug2ways: Reasoning over causal paths in biological networks for drug discovery. *PLOS Computational Biology*, 16(12): e1008464”.

Copyright © Rivas-Barragan D., *et al.*, 2020.

RESEARCH ARTICLE

Drug2ways: Reasoning over causal paths in biological networks for drug discovery

Daniel Rivas-Barragan^{1,2}, Sarah Mubeen^{3,4}, Francesc Guim Bernat⁵, Martin Hofmann-Apitius³, Daniel Domingo-Fernández^{3,4*}

1 Barcelona Supercomputing Center, Barcelona, Spain, 2 Computer Architecture Department, Universitat Politècnica de Catalunya, Barcelona, Spain, 3 Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing, Sankt Augustin, Germany, 4 Fraunhofer Center for Machine Learning, Germany, 5 Intel Corporation Iberia, Madrid, Spain

* daniel.domingo.fernandez@scai.fraunhofer.de



OPEN ACCESS

Citation: Rivas-Barragan D, Mubeen S, Guim Bernat F, Hofmann-Apitius M, Domingo-Fernández D (2020) Drug2ways: Reasoning over causal paths in biological networks for drug discovery. *PLoS Comput Biol* 16(12): e1008464. <https://doi.org/10.1371/journal.pcbi.1008464>

Editor: James R. Faeder, University of Pittsburgh, UNITED STATES

Received: July 5, 2020

Accepted: October 23, 2020

Published: December 2, 2020

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: <https://doi.org/10.1371/journal.pcbi.1008464>

Copyright: © 2020 Rivas-Barragan et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The data (i.e., networks) used in the case scenarios of the manuscript can be found at <https://github.com/>

Abstract

Elucidating the causal mechanisms responsible for disease can reveal potential therapeutic targets for pharmacological intervention and, accordingly, guide drug repositioning and discovery. In essence, the topology of a network can reveal the impact a drug candidate may have on a given biological state, leading the way for enhanced disease characterization and the design of advanced therapies. Network-based approaches, in particular, are highly suited for these purposes as they hold the capacity to identify the molecular mechanisms underlying disease. Here, we present drug2ways, a novel methodology that leverages multi-modal causal networks for predicting drug candidates. Drug2ways implements an efficient algorithm which reasons over causal paths in large-scale biological networks to propose drug candidates for a given disease. We validate our approach using clinical trial information and demonstrate how drug2ways can be used for multiple applications to identify: i) single-target drug candidates, ii) candidates with polypharmacological properties that can optimize multiple targets, and iii) candidates for combination therapy. Finally, we make drug2ways available to the scientific community as a Python package that enables conducting these applications on multiple standard network formats.

Author summary

At any given time, a large set of biomolecules are interacting in ways that give rise to the normal functioning of a cell. By representing biological interactions as networks, we can reconstruct the complex molecular mechanisms that govern the physiology of a cell. These networks can then be analyzed to understand where the system fails and how that can give rise to disease. Similarly, using computational methods, we can also enrich these networks with drugs, diseases and disease phenotypes to estimate how a drug, or a combination of drugs, would behave in a system and whether it can be used to treat or alleviate the symptoms of a disease. In this paper, we present drug2ways, a novel methodology designed for drug discovery applications, that exploits the information contained in a biological network comprising causal relations between drugs, proteins, and diseases.

[drug2ways/drug2ways](#) along with the code, all of which are openly available.

Funding: This work was developed in the Fraunhofer Cluster of Excellence "Cognitive Internet Technologies" (SM and DDF). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Employing these networks and an efficient algorithm, `drugways2` traverses over the ensemble of paths between a drug and a disease to propose the drugs that are most likely to cure the disease based on the information contained in the network. We hypothesize that this ensemble of paths could be used to simulate the mechanism of action of a drug and the directionality inferred through these paths could be used as a proxy to identify drug candidates. Through several experiments, we demonstrate how `drug2ways` can be used to find novel ways of using existing drugs, identify drug candidates, optimize treatments by targeting multiple disease phenotypes, and propose combination therapies. Owing to the generalizability of the algorithm and the accompanying software, we ambition that `drug2ways` could be applied to a variety of biological networks to generate new hypotheses for drug discovery and a better understanding of their mechanisms of action.

This is a *PLOS Computational Biology* Methods paper.

Introduction

Biological processes principally arise from interactions linking discrete biological entities. Far more rare, however, are processes that can be attributed to entities functioning in isolation. Hence, elucidating sets of interactions between biological entities is essential in understanding the mechanisms governing health and disease. Given the vast number of interactions that can occur in a particular biological system, these interactions are often abstracted and organized into large and highly interconnected computational networks. Many of the basic principles and methods from graph theory tend to be well-suited for network biology and applicable to various network types, such as protein-protein interaction (PPI), gene regulatory, and signaling networks [1]). Several discrete models, such as logical models [2] and Boolean networks [3,4] are common choices for their qualitative representation.

In a generic biological network representation, nodes denote entities, while edges denote their interactions. Multimodal networks can capture a wide range of biological scales, including physical entities (e.g., genes, proteins, and metabolites) or higher order concepts (e.g., biological processes, phenotypes, and diseases). Causal edges are those that possess directionality through direct interactions or through intermediaries [5]. These connections frequently occur in gene regulatory and metabolic/biochemical networks, while undirected edges are commonly present in chemical similarity or PPI networks to, for instance, represent symmetric binding relationships. For the latter group of edges, several methods [6,7] have emerged to assign directionality to interaction pairs (e.g., characterizing regulatory relationships as activation or inhibition relations) in order to assert causality which can be useful for various purposes. An example lies in discerning whether causal interactions between a drug target and intermediary proteins will inhibit a certain phenotype, a drug's intended effect, or activate it instead. Taken together, these networks enable a wide range of applications such as identifying disease mechanisms [8], making predictions on network perturbations [9], facilitating pathway analyses [10], establishing novel therapeutic drugs [11], and drug repurposing to detect potential therapeutic candidates [12].

Drug discovery is a major application that particularly benefits from network-based methods [11]. Typically, the traditional approach to drug discovery is characterized as follows: a drug target is selected based on an expressed phenotype, an assay is prepared for the target, high throughput screening (HTS) is performed, and hit or lead compounds are identified [13].

Though it may be the more conventional approach, the process tends to be laborious and is associated with both high costs and attrition rates. The latter can be attributed to several factors; firstly, experiments demonstrating the efficacy of drugs through their specific binding to a target may not be reproducible *in vivo* given the compartmentalization of the cell and/or the potential for other binding partners [14]. Secondly, in failing to investigate the cause of dysfunction that leads to disease within an appropriate biological context (e.g., molecular, cellular, or disease), the design of drugs is arbitrary [15]. These issues represent some of the prototypic problems that network-based approaches are ideally suited to address.

Beyond the utility of network-based methods for single target drug discovery and repurposing, these methods are also increasingly being used for the identification of pharmacological interventions that reverse multiple pathological states and in the design of drug combinations [16]. Although certain aspects of a pathology may be corrected by a single target drug, a multi-target drug or drug combination approach can have greater efficacy in reversing a disease or an expressed phenotype [17]. By taking into account causal mechanisms, network-based approaches can identify multiple targets within a network which, when modulated, can elicit synergistic effects [18]. Notably, combination therapies have successfully been used for several disease conditions including cancers [19,20] and the symptomatic management of Alzheimer's disease [21].

Various attributes of biological networks can serve as viable measures for network-based drug discovery. For instance, proximity measures such as the shortest path between a drug profile and a disease module have been used to identify potential drug repurposing candidates [22,23]. Additionally, centrality measures such as closeness and betweenness centrality also consider the shortest paths between pairs of nodes in order to pinpoint initial drug candidates [24,25]. However, potentially therapeutic targets may be connected to disease-relevant genes through paths not accounted for when solely considering shortest paths. Nevertheless, approaches which use non-shortest paths along a network are not without their limitations; as the size and complexity of networks increase, so too do the number of possible paths that can be traversed through the network, requiring greater computational power. Similarly, with an increasing number of nodes and edges, identifying multiple drugs for combination therapies that simultaneously target multiple disease-relevant genes and/or mitigate side-effects, can suffer from combinatorial explosions. Furthermore, not all paths in a network may be biologically plausible; erroneous interactions and those which are not biologically-relevant may also be present. Thus, making predictions for single and combination drug therapies can become highly challenging.

Here, we present drug2ways, a novel methodology applied to multimodal causal networks for the prediction of new drugs and the repurposing of existing ones. Our methodology consists of two main steps which jointly aim to address the high computational demands required to traverse large-scale, biological networks and to apply a reasoned approach to propose drug candidates for new indications by inferring causal paths. Firstly, drug2ways leverages a sophisticated and efficient algorithm to calculate all paths up to a given length between a drug and a disease or a set of phenotypes. Secondly, drug2ways traverses these paths to propose the set of drugs that are most likely to generate a desired phenotypic change. We demonstrate the utility of drug2ways for three different applications in order to identify: i) potential drug candidates, ii) potential candidates that optimize multiple target nodes of interest (i.e., indications and phenotypes) and iii) candidates for combination therapy. Finally, we make drug2ways available to the bioinformatics community as a Python package (<https://github.com/drug2ways>) that enables conducting the aforementioned applications on multiple standard network formats.

Results

We ambition multiple applications for drug2ways (Fig 1) which we present in three case scenarios and validate in two independent networks, the OpenBioLink knowledge graph (KG) and an In-House network. In the Subsection *Identifying drug candidates*, we first validate our methodology by showing how it can be used to identify potential drug candidates for various indications, while in the Subsection *Identifying drug candidates with multiple phenotypic targets*, we demonstrate how drug2ways can identify drugs that target sets of phenotypes present in specific indications. Finally, in the Subsection *Proposing combination therapies*, we show its utility in finding potentially efficacious drug combinations for combination therapy. In each of the three applications, the problem can be generalized to finding the relative effect of all paths between nodes representing chemicals and nodes representing phenotypes or clinical manifestations. Each application consists of reasoning over all possible paths of a predetermined length to evaluate the efficacy of either one or more chemicals in reverting the target node of interest (i.e., a manifestation and/or a set of associated phenotypes). This task can be conceived of as a brute-force search for all drugs and indications/phenotypes in a network for a given range of path lengths in order to prioritize drug candidates for each of the target nodes of interest.

The drug2ways algorithm incorporates two variants, namely all paths (i.e., a path in which repetition of vertices occurs) and simple paths (i.e., a path in which all vertices are distinct (and therefore, all edges)), enabling users to account for or ignore feedback loops (i.e., cycles), respectively (Fig 1D). Each of these three applications is associated with a high computational cost, especially the latter two which require calculations of a higher degree of complexity to identify potential candidates with multiple phenotypic/disease targets. However, because of the efficient implementation of the algorithm, each of these applications is attainable, which we demonstrate in the Subsection *Performance comparison and scalability of the algorithm* where we finally explore the scalability of drug2ways and compare it with standard path-finding implementations.

Identifying drug candidates

In Table 1, we summarize the results of drug2ways in recovering clinically-investigated drug-disease pairs for the top-ranked candidates in each of the validation experiments. Firstly, for both the original networks, drug2ways was able to retrieve a large proportion of drug-disease pairs that have been tested in clinical trials by calculating all paths up to a given length between a drug and an indication (i.e., l_{max}), although both networks exhibited differences based on the prioritization criteria described in the Subsection *Validation experiments*. For instance, the most restrictive prioritization criteria (i.e., 7/7 l_{max} inhibited the disease) yielded the best results for the In-House network, recovering nearly 40% of true positives from all prioritized pairs in the top-ranked list for all paths and simple paths respectively, while OpenBioLink yielded no prioritized pairs altogether. However, after a minimum relaxation of the prioritization criteria (i.e., 6/7 l_{max} inhibited the disease), OpenBioLink showed good results (i.e., ~50% and ~10% recovery rate for all paths and simple paths, respectively) while the recovery rate decreased for the In-House network to approximately 12%. In comparison, the proportion of true positives with respect to all possible combinations of drug-disease pairs from ClinicalTrials.gov is 3.19% for OpenBioLink (5.151/161.040) and 3.76% (9.537/253.638) for the In-House network, highlighting the significance of our results. These proportions are equivalent to the probability of randomly picking a true positive, which is comparatively much lower than the results yielded by drug2ways. In contrast, drug2ways failed to recover any true positives from the permuted versions of the original networks, further highlighting the validity of the results from the original networks.

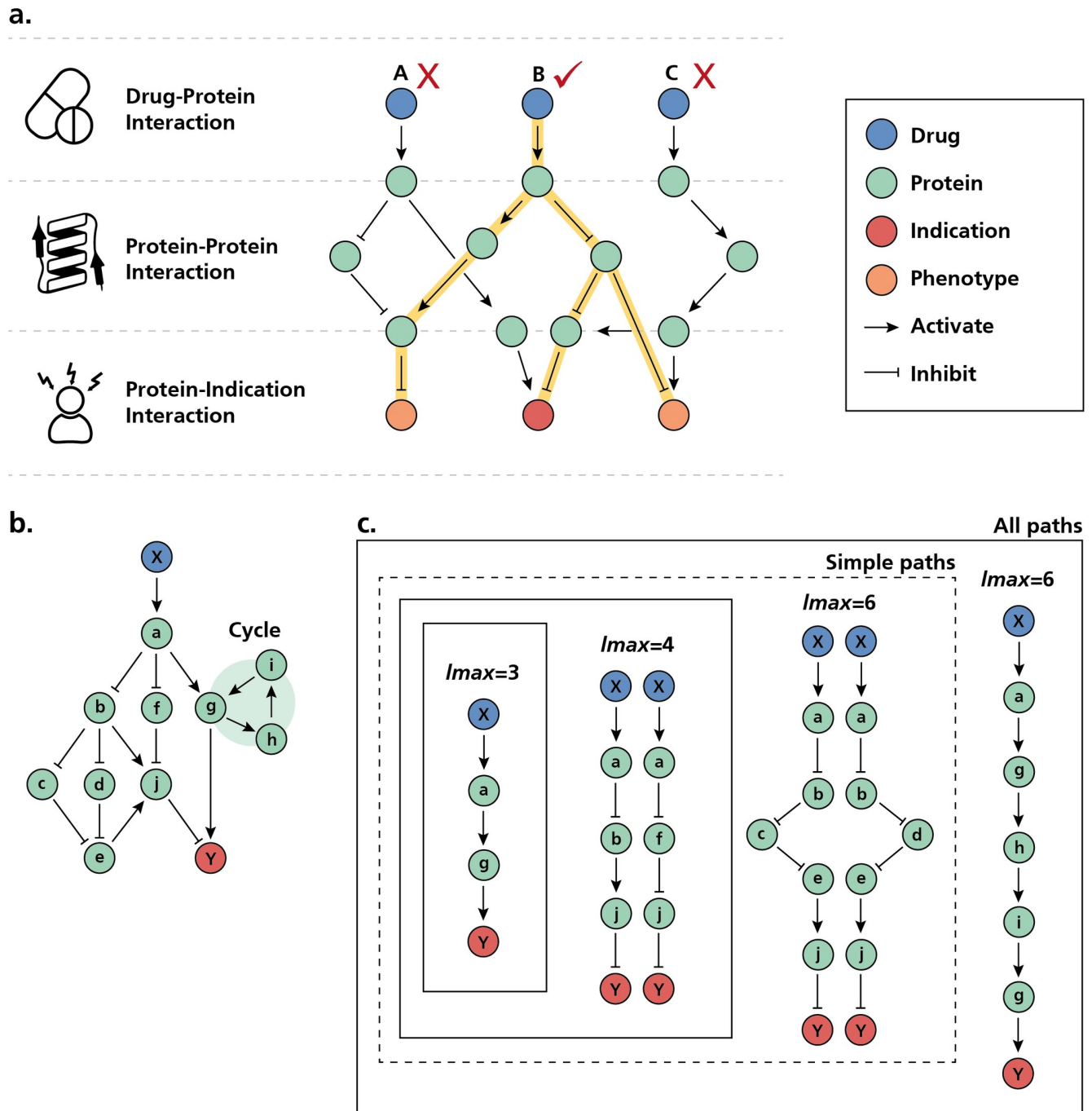


Fig 1. Schematic illustration of causal reasoning by drug2ways over simplified networks. a) Prototypic network used by drug2ways for drug discovery. The network contains causal relations between three modalities (i.e., drugs, proteins, and indications/phenotypes). Here, singular paths from three drugs to an indication as well as associated phenotypes are shown, though a single drug may contain multiple paths to a given indication/phenotype. Drug2ways reasons over all possible paths in a network between a drug and an indication/phenotype to predict the relative effect of each drug. In the example, we want to investigate whether one of the three drugs depicted inhibits an indication and its two phenotypes. While all three drugs target the disease, two of the three (i.e., drug A and C) fail to produce the desired effects (i.e., inhibition of the indication of interest and its two associated phenotypes). By reasoning over all the paths between the drug and the three target nodes of interest (i.e., indication and its phenotypes), drug2ways predicts that drug B could be a promising candidate as the majority of the paths would result in their inhibition, and thus produce a therapeutic effect. Similarly, drug2ways can also be used to evaluate the effect of a drug on a single indication/phenotype or to assess the effect of drug combinations. b) Example network containing all paths between a given drug and an indication. c) All possible paths between the drug and indication in (b). The drug2ways algorithm incorporates two variants, namely *all paths* and *simple paths*, enabling users to account for or ignore feedback loops (i.e., cycles), respectively. We distinguish between different paths based on the maximum number of allowable edges from a

drug X to an indication Y (i.e., $lmax$ parameter). For instance, the shortest path between the drug and the indication has an $lmax$ of 3 while an $lmax$ of 6 will capture this and four additional simple paths, two of length 4 and a further two of length 6. Using the *all paths* version of the algorithm, an additional cyclic path of length 6 is also captured.

<https://doi.org/10.1371/journal.pcbi.1008464.g001>

Given the small-world property of most biological networks, the predominant approaches in network-based drug discovery tend to investigate the shortest path between a drug and a disease. We thus compared our method to the shortest-path approach. While the results obtained using the shortest path are better than random, the shortest path tends to return a relatively high number of candidate pairs (>5.000) and a significantly lower recovery rate (~8%) than drug2ways (Table 1). Furthermore, we studied the lengths of the paths of candidate pairs prioritized by the shortest-path approach and, as expected, found that the vast majority of the paths are of lengths less than 4 (S1 Fig). In fact, the majority of the paths are $lmax = 2$, which corresponds to a direct drug-target-disease path. This indicates that the shortest-path approach can overlook diseases that are distant from drug targets, potentially explaining the difference in recovery rate between shortest-paths and drug2ways. Furthermore, while the shortest path only accounts for a single path between a drug and a disease, as an additional experiment, we investigated the total number of paths between all drug-disease pairs calculated from drug2ways using $lmax = 8$ to verify that predictions were not driven by the existence of a single path but by the directionality inferred through the ensemble of all paths (S2 Fig). We found that a large number of paths were present between most of the drug disease pairs, which when taken into account, could also explain the difference in the recovery rate.

The criteria selected for validation focused on prioritizing pairs exhibiting consistent scores (i.e., activation/inhibition ratio) through a wide range of $lmax$. In selecting this criteria, we intended to prevent any influence of path length (i.e., $lmax$) on the results. As expected, the results also indicate that the $lmax$ parameter and the prioritization criteria should be adapted for each new network. Thus, we recommend that users that intend to apply our methodology on their own networks follow a similar approach by using a broad range of $lmax$. Beyond the configuration of the $lmax$ parameter, we also recommend tuning a threshold value representing the relative effect of the drug on the indication, gradually decreasing this value to include additional, potential drug candidates. In this way, the Python implementation of drug2ways enables users to configure their experiments contingent upon the particular characteristics of the network (e.g., content and size).

Due to a lack of information on the directionality of protein-disease relations from high-quality resources, while generating both networks, we inferred association edges from DisGeNet [26] as activation edges (see Methods). Such a strong assumption implies that all proteins have an activation effect on the disease and ignores the possible inhibitory effects some of these proteins may have. Accordingly, due to this arbitrary inference, we hypothesized that

Table 1. Results of the validation experiments. The table presents the validation experiments for each of the four networks (i.e., OpenBioLink, permuted OpenBioLink, In-House, and permuted In-House) using two variants of the algorithm (i.e., all paths and simple paths) based on two different prioritization criteria (see Methods) as well as the results yielded when only considering the shortest path between a drug-disease pair. For each experiment, we report the relative number of true positives in the list of drug-disease pairs prioritized by drug2ways. The proportion of true positives recovered by both variants of drug2ways in the two original networks are significantly higher than chance level (i.e., 3.19% for OpenBioLink and 3.76% for the In-House network).

Network	All Paths		Simple Paths		Shortest Path
	7/7 Inhibit	6/7 Inhibit	7/7 Inhibit	6/7 Inhibit	
-					-
OpenBioLink	0/0 (0%)	2/4 (50%)	0/0 (0%)	1/11 (9.09%)	381/5.130 (7.43%)
Permuted OpenBioLink	0/0 (0%)	0/0 (0%)	0/0 (0%)	0/0 (0%)	40/5.130 (0.78%)
In-House	20/53 (37.74%)	105/919 (11.43%)	22/54 (40.74%)	106/872 (12.16%)	807/9.537 (8.46%)
Permuted In-House	0/0 (0%)	0/6 (0%)	0/0 (0%)	0/7 (0%)	274/9.537 (2.87%)

<https://doi.org/10.1371/journal.pcbi.1008464.t001>

some of the drug-disease pairs predicted as activating may indeed represent the opposite sign and also represent potential drug candidates. Thus, besides investigating drug-disease pairs that were consistently inhibited, we were also prompted to investigate pairs that were consistently activated. Confirming our hypothesis, we found that although based on our criteria, relatively few pairs were prioritized, clinically-investigated drug-disease pairs were also highly represented among the top-ranked active pairs (S3 Table).

In summary, our findings demonstrate the ability of drug2ways to recover a high proportion of clinically-tested drug-disease pairs. Due to our network design, candidate pairs consistently aggregate at both extremes of the distribution regardless of the relative directionality given by the ensemble of paths. Finally, among the novel drug-disease pairs that have not yet been tested in clinical trials, we have found multiple combinations reported in the literature, thus alluding to the potential for many other promising candidates for drug discovery that could be worth further exploring.

Identifying drug candidates with multiple phenotypic targets

The identification of drugs with several target nodes of interest (i.e., indications/phenotypes) can lead to more efficacious treatments, albeit their discovery is far more complex and thus represents a greater challenge than single-target drugs. In practice, this application is highly relevant as disease conditions can often manifest as sets of phenotypes. While the previous subsection demonstrated how our methodology is capable of identifying interesting single target drug candidates, in this subsection, we demonstrate how a network-based method can identify drug candidates that optimize multiple disease and phenotypic targets.

Here, we manually selected an indication and associated phenotypes present in both the In-House and OpenBioLink networks (S4 Table). Fig 2A illustrates the results of running the all

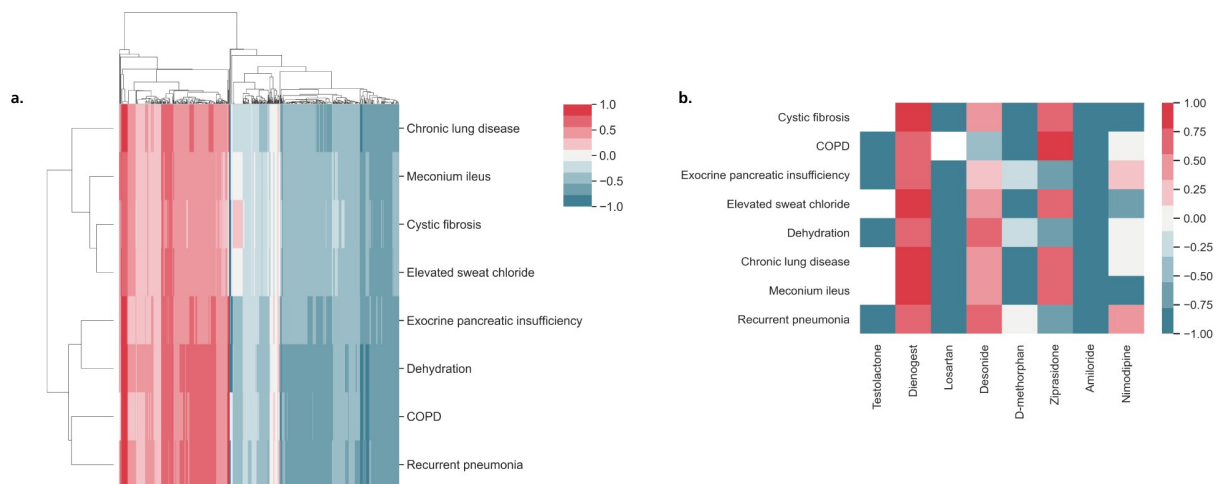


Fig 2. Identification of drugs targeting an indication and several associated phenotypes. The heatmaps summarize the results of running the all paths version of the drug2ways algorithm over the In-House network for variable path lengths. While the algorithm outputs scores between 0 and 1, where 0 denotes no activation or inhibition and 1 denotes a full activation or inhibition, scores were normalized between the range of -1 to 1. Here, normalized scores of the relative effects of drugs on cystic fibrosis and several of its associated phenotypes are displayed where values below and above 0 denote the inhibition (blue) and activation (red) of all paths between a drug and target indication/phenotype at a specific l_{max} , respectively, whilst 0 denotes a cancelling effect (gray). In a fourth case, no paths exist between the drug and indication/phenotype (white). **a)** Hierarchical clustering of normalized scores of the relative effects of all drugs in the In-House network on cystic fibrosis and related phenotypes at $l_{max} 8$. **b)** Heatmap illustrating a subset of drugs at $l_{max} 4$ which distinctly optimize therapeutic effects through inhibition of several disease/phenotypic targets (e.g., Amloride, D-methorphan, Losartan), activate the disease and/or its phenotypes (e.g., Dienogest), result in both the inhibition of some diseases/phenotypes and the activation of others (e.g., Desonide, Ziprasidone, Nimodipine), or do not possess paths to particular targets (e.g., Testolactone).

<https://doi.org/10.1371/journal.pcbi.1008464.g002>

paths version of the drug2ways algorithm over the In-House network at an $lmax$ of 8 for cystic fibrosis (CF) and seven related phenotypes. The heatmap shows that in selecting larger values of $lmax$, the vast majority of drugs (i.e., 626/671 drugs in the In-House network also in ClinicalTrials.gov) possess paths to each of the targets. We also note that most drugs in the network affect the indication and the phenotypes in a given direction (e.g., inhibition), while only a small minority will result in the activation of some phenotypes and/or indication and in the inhibition of others.

Once again, we altered the value of $lmax$ between 2 and 8 to investigate the relative effects of drugs yielded with varying path lengths. While beyond $lmax$ 4, we found little variation in the number of drugs containing paths to at least one target indication/phenotype (ranging from 602 drugs at $lmax$ 5 to 626 drugs at $lmax$ 8), we found fewer drugs at and below $lmax$ 4 (i.e., 55 at $lmax$ 2, 234 at $lmax$ 3, and 539 at $lmax$ 4). Fig 2B illustrates a subset of drugs at $lmax$ 4 that reverse, increase, cause no effect or have no paths to the indication and/or phenotypes. Among these drugs, we further investigated losartan, a drug under investigation in clinical trials for CF and studied the proteins implicated in paths of maximum length 4 between this drug and the disease. These proteins included *AGTR1*, whose reduced activity by pharmacological intervention has resulted in improved pulmonary functioning in mice with CF [27], and *TGFBI*, reduction of which by losartan has been shown to reverse mucociliary dysfunction related to inflammation and CF in animal models [28].

Proposing combination therapies

Combination therapies have been gaining major consideration for the treatment of disease and management of symptoms through the modulation of several targets by multiple drugs. However, with each additional drug for combination therapy, the task of identifying efficacious combinations by a network-based approach can result in a substantial increase in computational complexity, thus requiring efficient algorithms. Therefore, we were prompted to utilize drug2ways in a further application to explore the predicted effects of a combination of drugs on a given indication. We identified drug combinations consisting of pairs of drugs, though would like to note that our method could be used to identify combinations involving any number of drugs.

We manually selected several cancer types (i.e., breast cancer, colorectal cancer, lung cancer and melanoma) present in our In-House network to demonstrate an additional application of drug2ways to predict potential drugs for combination therapy. Similar to the previous two applications, as an input, we only considered drugs in the In-House network that were also present in ClinicalTrials.gov and used drug2ways to propose drug combinations at $lmax$ 4. For each of the four cancer subtypes, we then investigated existing drugs for their management and identified those that were also present in our network. We then focused on drug combinations that contained these drugs and caused inhibition of the cancer subtype. Table 2 lists a

Table 2. Examples of predicted combination therapies supported by literature evidence on four cancer types. The table reports drug combinations identified by drug2ways that inhibit each of the various cancer types and supporting literature evidence. These results were obtained by running the all paths version of the algorithm over the In-House network for $lmax$ 4.

Cancer type	Drug 1	Drug 2	Evidence
Breast cancer	Palbociclib	HCQ	[29]
Breast cancer	Palbociclib	Tamoxifen	[30]
Colorectal cancer	Palbociclib	Trametinib	[31]
Lung cancer	Dabrafenib	Trametinib	[32]
Lung cancer	Palbociclib	Trametinib	[33]
Melanoma	Mebendazole	Trametinib	[34]

<https://doi.org/10.1371/journal.pcbi.1008464.t002>

subset of drug combinations proposed by our methodology to inhibit specific cancer types and literature evidence on their potential therapeutic effects.

While, here we have only discussed drug combinations already in clinical trials or with correspondence to the literature, a multitude of combinations identified by our methodology that could potentially inhibit a disease but have not been reported thus far, represent potentially efficacious, novel combination therapies. Additionally, while in showcasing this functionality of our method, we have used all possible combinations of drugs that are both in our network and in clinical trials, this application can also be performed with a smaller set of drugs to evaluate the effect of particular drug combinations on a given set of diseases and/or phenotypes. Finally, each of the paths between a drug-disease pair can be defined as a sub-network representing biological processes and using pathway enrichment methods implemented in drug2ways, the mechanism of action of the drug can be elucidated.

Performance comparison and scalability of the algorithm

The applications described above have been conducted on large-scale networks comprising tens of thousands of nodes and edges, yet the size of biological networks can increase to incorporate millions. Therefore, the implementation of the algorithm has been designed to maximize its performance. Here, we compared drug2ways to the Python NetworkX library [35] (<https://networkx.github.io/>) and the C++/Python NetworkKit library [36] (<https://networkkit.github.io/>). We compare drug2ways against these two libraries, as both are widely used and already implement optimized methods for graph traversal and path retrieval. Both libraries implement a method to obtain all simple paths in a graph with a maximum path length. Fig 3 illustrates the runtime of each network-method pair in logarithmic scale on the y-axis, (i.e. for each network-method pair, the figure shows the time to count activation and inhibitory paths for each drug-disease pair in the network). As expected, the runtime is heavily dependent on the maximum path length l_{max} that we want to analyze. We added a timecap of 1.000 seconds (i.e. around 16 minutes) to the experiments, which is enough to show the method's scalability trendline and its exponential growth, while beyond this

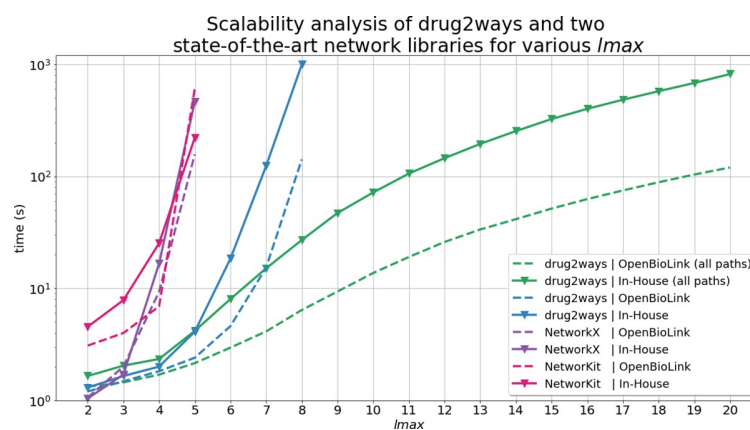


Fig 3. Average time required to calculate the effect of simple paths for all drug-disease pairs used in the validation on two heterogeneous networks using different l_{max} . The analysis was also conducted to take paths with repetitions of vertices between drug-disease pairs into account using the *all_paths* variant of drug2ways, but not for the NetworkX and NetworkKit libraries which lack equivalent implementations. Nevertheless, the implementations of both libraries could be easily adapted to return paths with repetitions of vertices. However, without the proper optimizations described in the Subsection *Theoretical background*, these would have a higher complexity than their *all_simple_paths* counterpart as nodes would be revisited. Therefore, for both libraries we use simple paths as the baseline for the analysis.

<https://doi.org/10.1371/journal.pcbi.1008464.g003>

timeframe, the runtime becomes unreasonably high. All three methods to count simple paths show a clear exponential growth in runtime. However, while NetworkX and NetworKit can be run with up to an $lmax$ of 5, drug2ways with simple paths is several orders of magnitude faster and is able to be run with up to an $lmax$ of 8. The comparison also shows that of the three different methodologies, only drug2ways can be scaled for large values of $lmax$ on both versions of the algorithm.

The *all_paths* variant of drug2ways does not show a pronounced exponential increase in time. However, the *all_simple_paths* variant shows a pronounced exponential increase in running time as it is computationally more expensive than *all_paths*. Here, the two standard libraries show a rapid exponential increase in time with $lmax$ values as low as 4 while drug2ways does not show a marked increase until values of $lmax$ beyond 7.

Taken together, we can see how the *all_paths* variant can be easily used for any large-scale network even when values of $lmax$ exceed 20, while the *all_simple_paths* variant requires both extensive computational power and time when such high values of $lmax$ are reached. In contrast, it is impractical to run experiments on large $lmax$ values using the other two standard libraries as they have not been optimized for the specific reasoning tasks presented in this work. Thus, these standard libraries would suffer from a high computation cost in conducting the applications of this approach (i.e., optimization of several phenotypes and/or an indication and identification of candidates for combination therapy), and in calculating paths on high values of $lmax$. Finally, we would like to note that in order to conduct a fair comparison, the experiments presented have not been conducted using the parallelization feature of drug2ways. Thus, we expect that in using this feature for the analysis, the difference in the performance between drug2ways and the other two libraries would have been even more pronounced.

Discussion

Increasingly, network-based methods are emerging as promising alternatives to traditional approaches for drug discovery by taking into account causal mechanisms responsible for disease. Here, we have presented a robust and efficient method that leverages causal interactions in biological networks to predict drug candidates for a given disease or a set of phenotypes, as well as pairs of drugs for combination therapy. While previous methods have focused on leveraging network proximity methods (e.g., shortest paths) between drugs and indications [22,23], drug2ways leverages all the paths between a given drug and disease. Although not all paths in a network may be plausible as some paths may be irrelevant or erroneous, we hypothesize that by reasoning over a multitude of possible paths, we can estimate the relative effect of each drug on a disease as the average of all possible paths. In doing so, we assume that a drug has a greater likelihood of modulating a disease as the number of possible paths connecting a drug to a disease increases. Therefore, exploring all paths in which a drug could modulate a disease or a phenotype can serve as a proxy for the prediction of novel drugs. To test our hypothesis, we systematically predicted the effect of each drug on all diseases in two multimodal networks of different size and content. Next, we validated our results against clinical trial information showing that our approach could retrieve a large proportion of true positives. Furthermore, with a second application, we demonstrate the ability of our approach to identify single drugs that can simultaneously modulate multiple targets to revert a set of phenotypes. Finally, the third application shows how a similar strategy can be applied for combination therapy.

Although drug2ways requires multimodal networks that contain causal relations between drugs, proteins, and indications/phenotypes, it can also be tuned and applied to other networks with different properties. For instance, we propose the use of networks comprising non-molecular nodes, such as biological processes, in cases when molecular information is not

widely available. Given the exponential increase in computational complexity when using the algorithm on multiple drugs for combination therapy, we demonstrate this application exclusively on drug pairs. Nonetheless, the high performance of drug2ways, which also allows for parallelization, enables users to conduct experiments upon millions of combinations in contrast to other state-of-the-art network libraries which would require an immense amount of time and computational resources.

One of the major limitations of this work is the absence of signed causal information regarding the effect of proteins on indications and phenotypes. To circumvent this issue, we inferred all protein-indication and protein-phenotype associations as activations, an assumption that may not correspond to the true biology. Thus, due to a lack of such information, curating and qualifying directionality for these relations could be a future improvement for drug2ways. Additionally, we would like to acknowledge the possible effects of feedforward loops on the results, especially as l_{max} increases. However, the design of our validation has taken this factor into consideration. Finally, although we validated our results with clinical trial information and tested the robustness of our approach, by simplifying biology to a network of binary causal relationships, we overlook its quantitative aspects. Therefore, we would like to note that quantitative measures, such as kinetic rates for reactions, the confidence of the interaction, and the magnitude of the effect, may provide a more realistic representation and thus, could be considered in future work by adding these aspects as weights to the edges in a network. Finally, we also intend to investigate the feasibility of drug2ways to identify drugs that mimic disease phenotypes and hence, could be potentially employed to generate *in vitro* or *in vivo* models.

In summary, our approach demonstrates that reasoning over multiple causal paths in biological networks can potentially serve to predict candidates for drug discovery. From a translational perspective, drug2ways can be used to identify novel drugs and combination therapies for indications where their mechanisms of action can be well represented in a network. Finally, we provide a user-friendly Python package that enables conducting the three presented applications on biological networks in multiple standard formats.

Methods

In the first four subsections, we outline relevant graph theoretical concepts, describe the graph traversal algorithm presented in the study, delineate its complexity, and provide details on the implementation of the software. Next, we discuss applications of the algorithm which are illustrated in case scenarios and validation experiments. In the final subsection, we provide details on the hardware used.

Theoretical background

Given that most biological networks display the small-world property in which paths between pairs of nodes are relatively short, many genes can be in the vicinity of disease-relevant ones [14]. Accordingly, a simple yet effective approach to identifying potential drug targets is to consider nodes that are in close proximity to disease genes. However, not all of these nodes may necessarily be linked to disease genes, but rather, may simply be false positives resulting from spurious or irrelevant interactions [37]. Furthermore, such an approach can overlook interesting genes linked to disease-relevant ones by longer, alternative paths. One possible solution to this problem lies in traversing all possible paths between a pair of nodes to reach beyond the limits of local, proximity-based approaches. Beyond calculation of all paths between a drug and disease-related gene, however, a reasoned approach can be used to suggest how a drug may modulate a disease given the number of paths and types of interactions

between the two. Essentially, with a causal network containing directed relationships, signed -1 to indicate inhibition and +1 to indicate activation, we can define the relative effect of each drug as the proportion of activatory/inhibitory paths from all possible paths between the two (Fig 1). Nonetheless, with several thousand drugs and diseases, the computational complexity to traverse all possible paths between each pairwise combination can increase dramatically.

An intuitive solution to determine the relative effect of a drug on an indication would be to first find the set of all paths between them and then compute the effect on each of these paths. However, the problem of finding all paths in a network, which we will interchangeably refer to as a graph, is known to be NP-Hard (i.e., computationally hard), which are the class of problems in computational complexity that are not solvable in polynomial time. This makes the problem intractable as with an increasing number of vertices for some types of graphs (e.g., fully connected graphs), the total number of paths grows exponentially. However, to solve this problem we are not required to store the whole sequence of edges forming each path. Instead, if edges in a path are represented by their effects (i.e., -1 and +1 labels indicating inhibition and activation, respectively), we can define the combined effect of the path as the product of all edges it contains, while for the same set of edges regardless of the order they appear in the graph, the combined effect will always remain the same. This enables a series of optimizations which allow us to reduce time and space complexity, as explained in detail in the Subsection *Algorithm*. If a graph contains cycles (i.e., feedback loops), an infinite number of possible paths can be found by repeating the sequence of edges containing the cycle (Fig 1B). However, an increasing number of possible edges can also lead to an exponential increase in the number of paths, most of which may not be biologically plausible and result in the true biological effect becoming lost. We thus consider paths only up to a maximum length to limit the influence of cycles and highly elongated paths whilst still capturing feedback loops (Fig 1C).

We first define a series of terms that will be used throughout this section to provide a formal definition of the problem. Given an unweighted directed graph $G = (V, E)$, V is the set of vertices (interchangeably nodes) and E is the set of edges in the graph. A path is defined as a sequence of edges (e_1, e_2, \dots, e_k) that joins a sequence of vertices $(v_1, v_2, \dots, v_{k+1})$ in a graph, for $1 \leq k \leq |E|$ such that $e_i = \{v_i, v_{i+1}\}$, for $1 \leq i \leq k$, where k is the number of edges and the length of the path. Consequently, we denote a path between a source node s and a target node t as $p_{s,t}$ for $s, t \in V$ i.e. for the set of nodes $(v_1, v_2, \dots, v_{k+1})$ joined by the path, $s = v_1$ and $t = v_{k+1}$, while nodes v_i , for $1 \leq i \leq k+1$ are *intermediate nodes* (see Table 3 for key definitions). Similarly, a *cyclic path* is a path when the first and last vertices it joins are the same, while a *simple path* is a path where all vertices are distinct. Furthermore, any edge $e \in E$ in G represents a relationship between the pair of nodes it connects and it is labeled +1 or -1 depending on whether it is an activatory or an inhibitory relationship, respectively. Following, the effect that a node $s \in V$ has on node $t \in V$ over a given path $p_{s,t}$ is computed as $effect(p_{s,t}) = \prod_{i=1}^k e_i$, $\forall e \in p_{s,t}$, where $e_i \in \{-1, +1\}$ and is the label

Table 3. Definitions of terms used in this paper.

Term	Definition
Simple path	A path in which all vertices are distinct (and therefore, all edges).
Cyclic path	A path in which repetition of vertices occurs.
All paths	The set of all paths, including those which contain cycles.
Intermediate node	Any node v in a path between two nodes u, t , s.t. $v \notin \{u, t\}$.
Path length	The number of edges in a path between a source node and a target node.
l_{max}	The maximum length of the paths between a source and target node. In other words, for any given l_{max} , only paths with a length less than or equal to l_{max} are considered.

<https://doi.org/10.1371/journal.pcbi.1008464.t003>

of the i^{th} edge in the path. A path $p_{s,t}$ is said to be an activatory path if its effect is equal to +1. Analogously, the path is said to be an inhibitory path if its effect is equal to -1.

Before defining the problem, we would like to remark that $p_{s,t}$ does not necessarily represent a singular path; as s and t might be connected by multiple sets of edges and different sets of edges may yield different effects between the nodes, a path is uniquely identified if its entire sequence of edges is unique. Furthermore, we would also like to remark that once the effect of a path is computed, we are no longer interested in the set of edges and intermediate nodes of a given path. Therefore, for simplicity, we define $P_{s,t}$ as the set of all paths between s and t . Similarly, $A_{s,t}$ denotes the set of all activatory paths between s and t and $I_{s,t}$ the set of all inhibitory paths between s and t .

Finally, we define the problem as follows: given an unweighted directed graph $G = (V, E)$, a subset of vertices $D \subset V$, representing drugs, and a subset $T \subset V$, representing target phenotypes, we are interested in finding the relative effect of a node s over a node t $\delta_k(s, t) = \frac{|\epsilon_{s,t}|}{|P_{s,t}|}$ for $s \in D$, $t \in T$ and $1 \leq k \leq |E|$, where $\epsilon_{s,t}$ is equal to $A_{s,t}$ or $I_{s,t}$, depending on the effect we are interested in. For instance, if we want to investigate whether a drug could reverse a phenotype, we would compute the proportion of inhibitory paths over all paths of length less than or equal to k between the pair of nodes.

Algorithm

From the previous definition of relative effect, (e.g., $\delta_k(s, t) = \frac{|I_{s,t}|}{|P_{s,t}|}$ for the relative inhibition), its computation requires that activatory and inhibitory paths between nodes s and t are counted independently. The number of paths from s to t with length less than or equal to k can be defined as the sequence shown in Eq 1. From the equation, it is intuitive to think of a recursive implementation to traverse the graph using a modified version of the DFS (Depth First Search) algorithm. This definition yields the foundations for an intuitive yet optimized algorithm by means of *dynamic programming* and *memoization*.

$$all_paths(s, t, k) = \{1 \text{ if } s = t, \text{ otherwise } \sum path\ s(u, t, k - 1) \forall u \in neighbors(s)\} \quad \text{Eq 1}$$

Dynamic programming is a method for solving a complex problem by breaking it down into simpler problems whose solutions are part of the former's solution. From Eq 1, we can easily extract that the problem of finding the number of paths from s to t can be broken down to finding the number of paths from all neighbors of s to t , with maximum length of $k-1$. Once a solution for $all_paths(u, t, k)$ is found, for any $u \in V$, it is stored and used whenever it is a sub-problem to be solved again. This optimization technique is called *memoization* and is what guarantees that a node is never revisited with the same length k .

We have implemented two variants of drug2ways to calculate the relative effect of a pair of nodes, namely *all_paths* and *all_simple_paths* (detailed explanation and pseudocode in the [S1 Text](#)). The former considers all paths between two nodes in the graph, i.e. including *cyclic paths*, while the latter considers only *simple paths* (Table 3). This differentiation is important because *all_simple_paths* adds the restriction that cycles must be avoided and with it comes a higher complexity of the algorithm, as some nodes might be revisited. In order to evaluate the scalability of our methodology with respect to comparable methods for graph traversal, in the Subsection *Performance comparison and scalability of the algorithm*, we analyzed the performance of two variants of drug2ways (i.e., *all_paths* and *all_simple_paths*) to obtain the number of activating and inhibiting paths between pairs of nodes. We then compared the performance of drug2ways against two equivalent path-finding methods implemented in two state-of-the-art network libraries.

Complexity

Both variants of `drugs2ways` (i.e. `all_paths` and `all_simple_paths`) traverse the graph visiting nodes recursively in DFS order with a maximum path length k . However, as previously stated, reasoning over all paths versus only simple paths are two different problems with disparate computational complexities. In the first variant of `drug2ways` (i.e., `all_paths`), once a node is visited, it is never revisited for a path length less than or equal to k , as the intermediate result stored in the cache is enough to guarantee a valid solution (S1 Text: Algorithm 1). In the worst case, a node is visited l_{max} times, with l_{max} being the maximum path length when the algorithm starts. Therefore, `all_paths` has a complexity of $O(l_{max} \times |V|)$. As for space, the cache stores two integer values (activatory and inhibitory paths are counted separately) for each pair of nodes u, t for $u \in T^C \subset V$ and $t \in T \subset V$ and for each length k $1 \leq k \leq |E|$, for which a node has been visited. This translates to an upper bound in space of $\frac{|V|^2}{4}$. Thus, the algorithm has a space complexity of $O(|V|^2)$. Nevertheless, we would like to note that for biological graphs and the applications of the algorithm we devised, it is rarely the case that every target node in the graph is explored. As a consequence, the complexity is lower on the average case, as the number of target nodes is usually a small subset of V and the number of targets to explore is in the order of units. On the other hand, the second variant of `drug2ways` (i.e., `all_simple_paths`) revisits a node every time a cycle is detected (S1 Text: Algorithm 2). This increases the complexity to $O(|V|^{l_{max}})$ in the worst case. However, the average case is still several orders of magnitude faster than other standard algorithms, as discussed in the Subsection [Performance comparison and scalability of the algorithm](#).

Software and implementation

To facilitate the usage of the algorithm presented in the previous section, we implemented it in a Python package called `drug2ways`. The package leverages state-of-the-art Python packages such as `NetworkX` for network analysis [35], `MPI` for parallelization (<https://mpi4py.readthedocs.io/>), and `click` for exposing the command line interface (CLI) (<https://click.palletsprojects.com>). `Drug2ways` allows users to use the algorithm on a variety of standard network formats (e.g., `GraphML`, `Node-Link`, and `EdgeList`) and is powered by a CLI, following the standard proposed by [38]. The CLI offers all the case scenarios for proposing drug candidates that are presented in the results section.

The Python package is available at <https://github.com/drug2ways/drug2ways>, its latest documentation can be found at <https://drug2ways.readthedocs.io> and its distributions can be found on PyPI at <https://pypi.org/project/drug2ways>. Finally, the scripts for generating the figures in this manuscript are included in Jupyter notebooks at <https://github.com/drug2ways/results>.

Case scenarios

Networks. To demonstrate the above-mentioned applications, we used two different multimodal networks of varying size and content (Fig 4). Although each of the two networks contain unique interactions depending on the source databases they include, both the networks incorporate the following types of relations: drug-protein, protein-protein, protein-indication, and protein-phenotype. Minimally, we required each of these relationships as they simulate the binding of a drug to a target (i.e., drug-protein relation), the triggering of a cascade of events (i.e., a set of protein-protein interactions), and an effect on an indication or a phenotypic observation (i.e., protein-phenotype/indication associations), respectively. Notably, while all relationships maintained their original directionality from their source database, protein-

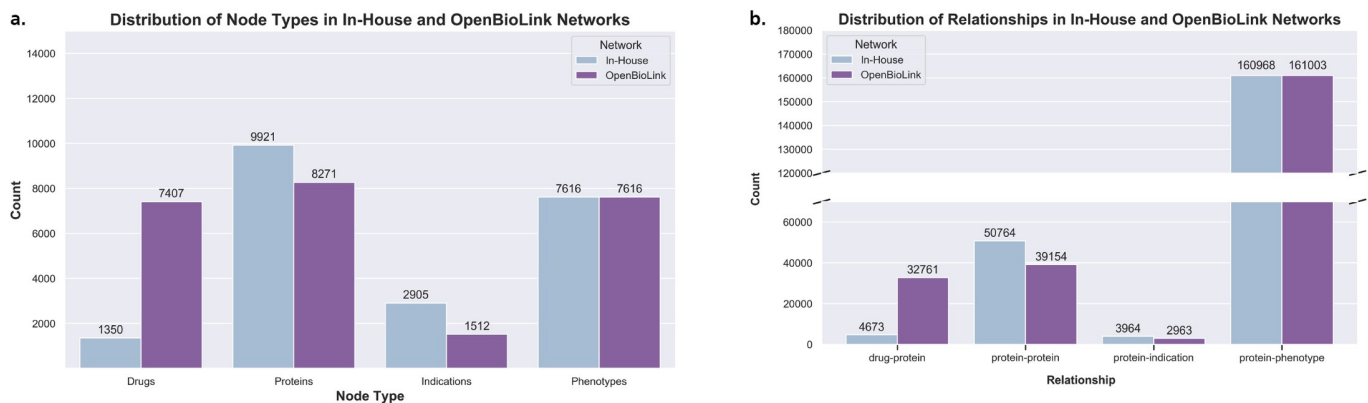


Fig 4. Distribution of node types and relationships in the In-House and OpenBioLink networks. a) The OpenBioLink KG contains a greater proportion of PubChem drugs relative to the In-House network which solely contains drugs from DrugBank. While the number of proteins in each of the two networks is comparable, indications are more numerous in the In-House network with respect to the OpenBioLink KG. Phenotypes for the In-House network were sourced from OpenBioLink, and as such, are equivalent in number. b) The total number of drug-protein interactions is greater in the OpenBioLink network than in our In-House. A greater proportion of protein-protein interactions are present in the In-House network, as are the number of protein-indication edges while the number of protein-phenotype interactions are nearly equivalent.

<https://doi.org/10.1371/journal.pcbi.1008464.g004>

phenotype and protein-indication associations lacked explicit causal information and were thus inferred as activation relationships. Details about the types of each interaction are provided in S1 and S2 Tables. Below, we describe each of the two networks used.

The first network, OpenBioLink, is a large-scale KG generated from an integrative effort designed to establish a benchmark dataset for link prediction [39]. The second is an In-House network that is comprised of tens of thousands of interactions from eight databases that we have harmonized for this work including PathMe [40–43], BioGrid [44], IntAct [45], and PathwayCommons [46] for protein-protein relations, DrugBank [47] for drug-protein relations, and DisGeNet [26] for protein-indication interactions. In addition to these eight databases, protein-phenotype relationships were sourced from the OpenBioLink KG.

Validation experiments. In the first of three validation experiments, we ran the algorithm on two versions (all paths vs simple paths) of each of the networks over a wide range of l_{max} . We selected 2 as the minimum l_{max} as we require at least one intermediate target node between a drug and an indication. In choosing 2 as the lower bound, we incorporate the shortest possible path between a drug and an indication. However, our approach was focused more heavily on elaborate paths as a means to exploit a greater degree of complexity in biological networks. Accordingly, we set 8 as an upper bound for l_{max} such that longer paths connecting a target and a disease could also be explored. Above this range, the score, defined as the proportion of activatory/inhibitory paths (i.e., activation/inhibition ratio) tends to converge as the effect of a drug appears to cancel itself out through several, contradictory interactions (S2 Text and S1 Appendix). This event is altogether unsurprising and could be partially explained by interactions that may not be biologically plausible and through the exploration of distant nodes. Thus, users that intend to use our methodology on a different network should first study the distribution of scores as l_{max} increases, prior to determining an optimal l_{max} range. The reason is that an optimal l_{max} range can vary depending on the characteristics of a network (e.g., size, number of activation versus inhibition interactions, average number of connections, etc). Finally, we would also like to mention that a significant increase in computational time would be required for the algorithm to run for larger values of l_{max} as the number of paths with an l_{max} of 8 exceeds several millions for numerous drug-disease pairs (see Subsection *Performance comparison and scalability of the algorithm*).

Table 4. Clinical trial information mapped to the OpenBioLink and In-House networks for drug2ways validation. The procedure to extract the information from ClinicalTrials.gov and the corresponding lists of drugs and diseases are available at <https://github.com/drug2ways/results/tree/master/validation>.

Network	Drug-Disease Pairs from ClinicalTrials.gov	Unique Drugs	Unique Diseases	Possible Combinations
OpenBioLink	5.151	610	264	161.040
In-House Network	9.537	671	378	253.638

<https://doi.org/10.1371/journal.pcbi.1008464.t004>

In the second experiment, we sought to validate drugs which could be effective against a given disease by incorporating clinical trial information in line with similar recent validation approaches in the literature [48,49]. As clinical trial investigations evaluate the effects of drug interventions for various indications, drug-disease pairs from ClinicalTrials.gov were used as the ground-truth list of positive labels. In total, 59,798 unique drug-disease pairs were extracted from the ClinicalTrials.gov website on 16-04-2020. Since our approach will only find paths between pairs when both the drug and disease are present in the network, only those pairs from ClinicalTrials.gov that could map to OpenBioLink and the In-House network were used as positive labels (Table 4). Thus, the original list of 59,798 unique drug-disease pairs was reduced according to the number of pairs that could be mapped to each network (i.e., 5,151 for OpenBioLink, and 9,537 for the In-House network). To conduct the validation experiments, we ran drug2ways using the drugs (source nodes) and the diseases (target nodes) present in these two filtered lists of positive labels, corresponding to a total of 161,040 possible pairs for OpenBioLink and 253,638 for the In-House network (Table 4).

Our approach exhibits the so-called early retrieval problem, or in other words, from the thousands of possible combinations of drug-disease pairs, only the top-ranked pairs contain interesting candidates for drug discovery. For such classification tasks, conventional metrics such as receiver operating characteristic (ROC) curves (i.e., AUC-ROC and AUC-PR) become inadequate [50]. This is because a classifier may accurately predict positive cases in the top-ranked pairs, but have a low predictive performance in the remaining cases that are not particularly interesting for drug discovery, leading to Area Under the Curve (AUC) values close to 0.5. For example, imagine a scenario in which 150,000 combinations of drug-disease pairs are possible in the OpenBioLink network, and of these, 5,000 are positive labels (i.e., 3%). From all possible combinations, if we consider the top 100 pairs prioritized by drug2ways and of these, 50 (i.e., 50%) are true positives, then drug2ways has captured a significantly greater number of true positives (50%) than what is expected by chance (3%). However, depending on the ranking of these pairs, it is possible to obtain a low AUC-ROC if the true positives are fairly distributed across this list of 100 pairs. Furthermore, some of these prioritized pairs may represent potential drug-disease pairs that have not been investigated before. Finally, we would like to note that only 3% of drug-disease pairs are positive labels in both networks; thus, implying a significant imbalance of class labels (Table 4). In light of these shortcomings, we have evaluated drug2ways using the AUC-ROC as a metric, yielding an AUC value of approximately 0.65 for both networks and versions of the algorithm (S3 and S4 Figs). Nonetheless, we also present a validation based on the ratio of true positives that appear in the top-ranked drug-disease pairs in order to evaluate the top-ranked set of pairs prioritized by drug2ways. Subsequently, we prioritized these pairs if they fulfilled the prioritization criteria as follows (see examples in Table 5):

1. **High inhibition.** Since we are interested in identifying drugs that inhibit a particular indication, for a pair to be prioritized, we required that at least 75% of the paths between the pair must be predicted to inhibit the indication. As we empirically selected this value, we also studied the effect of this parameter on the performance of drug2ways in the S5 and S6 Tables.

Table 5. Illustration of the prioritization with three example pairs (i.e., A, B, and C). For each l_{max} , the number and percentage of inhibitory paths is shown. While all three pairs show a similar pattern, pair B has less than 70% of inhibitory paths for $l_{max} = 3$ (i.e., Criterion 2) while for pair C, an increase in the number of paths from $l_{max} = 2$ to $l_{max} = 3$ does not occur (i.e., Criterion 3). Finally, pair A fulfills all three criteria and can thus be categorized as a prioritized pair.

Pair	l_{max}							Prioritize
	2	3	4	5	6	7	8	
A	1 (80%)	4 (90%)	20 (100%)	50 (100%)	100 (80%)	400 (90%)	1.000 (80%)	Yes
B	1 (80%)	4 (70%)	20 (100%)	50 (100%)	100 (80%)	400 (90%)	1.000 (80%)	No
C	1 (80%)	1 (90%)	20 (100%)	50 (100%)	100 (80%)	400 (90%)	1.000 (80%)	No

<https://doi.org/10.1371/journal.pcbi.1008464.t005>

- Consistent inhibition.** The second criteria aimed at testing the stability of the predicted effect for a given pair independent of changes to l_{max} . Accordingly, we only consider pairs where the previous criteria (i.e., more than 75% of the paths inhibit the disease) is maintained through the l_{max} range used (i.e., from 2 to 8).
- Increasing number of paths.** With each incremental increase in l_{max} , the number of paths must also increase such that novel paths are reported at every step of l_{max} .

As a third and final validation, we compared the two prioritized lists for each network against random lists generated by permuted versions of the original networks that were created using the XSwap algorithm [51]. By using this algorithm, we ensured that the permuted versions preserved the original structure of the original network (i.e., each node has the same number of in- and out-edges) as well as maintained the same number of activation and inhibition edges.

Hardware

Computations for each of the tasks were performed on a symmetric multiprocessing (SMP) node with four Intel Xeon Platinum 8160 processors per node with 24 cores/48 threads each (96 cores/192 threads per node in total) and 2.1GHz base / 3.7 GHz Turbo Frequency with 1536GB/1.5TB RAM (DDR4 ECC Reg). The network was 100Gbit/s Intel OmniPath, storage was 2x Intel P4600 1.6TB U.2 PCIe NVMe for local intermediate data and BeeGFS parallel file system for Home directories.

Supporting information

S1 Fig. Frequencies of the lengths of the shortest-paths calculated between all drug-disease pairs with $l_{max} \leq 8$ in the OpenBiolink and In-House networks.

(DOCX)

S2 Fig. Distribution of total paths between all drug-disease pairs in the OpenBiolink and In-House networks with $l_{max} = 8$.

(DOCX)

S3 Fig. The AUROC curves for both networks presented in the case scenario using the all paths version of drug2ways.

(DOCX)

S4 Fig. The AUROC curves for both networks presented in the case scenario using the simple paths version of drug2ways.

(DOCX)

S1 Table. Relationships in the In-House network and their assigned polarity.

(DOCX)

S2 Table. Relationships in OpenBioLink and their assigned polarity.

(DOCX)

S3 Table. Results of the validation experiments focusing on prioritized drugs that activate an indication.

(DOCX)

S4 Table. Phenotypes associated with cystic fibrosis of pancreas, the indication investigated in the Subsection *Identifying drug candidates with multiple phenotypic targets*.

(DOCX)

S5 Table. Effect of the percentage of inhibitory paths on the number of true positives (6/7 l_{max} inhibit).

(DOCX)

S6 Table. Effect of the percentage of inhibitory paths on the number of true positives (7/7 l_{max} inhibit).

(DOCX)

S1 Appendix. “score_distributions.zip”. Distribution of the scores for each l_{max} value on both networks

(ZIP)

S1 Text. Algorithm.

(DOCX)

S2 Text. Comparing distribution scores between the original and permuted networks.

(DOCX)

Acknowledgments

The authors would like to thank Sophia Krix for her assistance generating the networks and Colin Birkenbihl for his valuable feedback.

Author Contributions

Conceptualization: Daniel Domingo-Fernández.

Data curation: Sarah Mubeen, Daniel Domingo-Fernández.

Formal analysis: Daniel Rivas-Barragan, Sarah Mubeen, Daniel Domingo-Fernández.

Funding acquisition: Francesc Guim Bernat, Martin Hofmann-Apitius, Daniel Domingo-Fernández.

Investigation: Daniel Rivas-Barragan, Sarah Mubeen, Daniel Domingo-Fernández.

Methodology: Daniel Rivas-Barragan, Daniel Domingo-Fernández.

Project administration: Francesc Guim Bernat, Martin Hofmann-Apitius, Daniel Domingo-Fernández.

Resources: Daniel Rivas-Barragan, Sarah Mubeen, Daniel Domingo-Fernández.

Software: Daniel Rivas-Barragan, Daniel Domingo-Fernández.

Supervision: Daniel Domingo-Fernández.

Validation: Daniel Rivas-Barragan, Sarah Mubeen, Daniel Domingo-Fernández.

Visualization: Daniel Rivas-Barragan, Sarah Mubeen, Daniel Domingo-Fernández.

Writing – original draft: Daniel Rivas-Barragan, Sarah Mubeen, Daniel Domingo-Fernández.

Writing – review & editing: Daniel Rivas-Barragan, Sarah Mubeen, Daniel Domingo-Fernández.

References

1. Pavlopoulos GA, Secrier M, Moschopoulos CN, Soldatos TG, Kossida S, Aerts J, et al. Using graph theory to analyze biological networks. *BioData mining*. 2011; 4 (1):10. <https://doi.org/10.1186/1756-0381-4-10> PMID: 21527005
2. Naldi A, Thieffry D, Chaouiya C. Decision diagrams for the representation and analysis of logical models of genetic networks. *Proceedings of the 2007 international conference on International Conference on Computational Methods in Systems Biology*. Berlin, Heidelberg: Springer-Verlag; 2007:233–247. 2007. https://doi.org/10.1007/978-3-540-75140-3_16.
3. Kauffman SA. Metabolic stability and epigenesis in randomly constructed genetic nets. *J Theor Biol*. 1969; 22 (3):437–67. [https://doi.org/10.1016/0022-5193\(69\)90015-0](https://doi.org/10.1016/0022-5193(69)90015-0) PMID: 5803332
4. Shmulevich I, Dougherty ER, Kim S, Zhang W. Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*. 2002; 18 (2):261–74. <https://doi.org/10.1093/bioinformatics/18.2.261> PMID: 11847074
5. Hill SM, Heiser LM, Cokelaer T, Unger M, Nesser NK, Carlin DE, et al. Inferring causal molecular networks: empirical assessment through a community-based effort. *Nat Methods*. 2016; 13 (4):310–8. <https://doi.org/10.1038/nmeth.3773> PMID: 26901648
6. Wang T, Feng Y, and Wang Q. (2017). PAIRS: Prediction of Activation/Inhibition Regulation Signaling Pathway. *Computational intelligence and neuroscience, 2017*. <https://doi.org/10.1155/2017/7024516> PMID: 28469669
7. Yim S, Yu H, Jang D, Lee D. Annotating activation/inhibition relationships to protein-protein interactions using gene ontology relations. *BMC Syst Biol*. 2018; 12 (1):9. <https://doi.org/10.1186/s12918-018-0535-4> PMID: 29671402
8. Saqi M, Lysenko A, Guo YK, Tsunoda T, Auffray C. Navigating the disease landscape: knowledge representations for contextualizing molecular signatures. *Brief Bioinform*. 2019; 20 (2):609–23. <https://doi.org/10.1093/bib/bby025> PMID: 29684165
9. Santolini M, Barabási AL. Predicting perturbation patterns from the topology of biological networks. *Proc Natl Acad Sci*. 2018; 115 (27):E6375–83. <https://doi.org/10.1073/pnas.1720589115> PMID: 29925605
10. Reimand J, Isserlin R, Voisin V, Kucera M, Tannus-Lopes C, Rostamianfar A, et al. Pathway enrichment analysis and visualization of omics data using g: Profiler, GSEA, Cytoscape and EnrichmentMap. *Nat Protoc*. 2019; 14 (2):482–517. <https://doi.org/10.1038/s41596-018-0103-9> PMID: 30664679
11. Csermely P, Korcsmáros T, Kiss HJ, London G, Nussinov R. Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. *Pharmacol Ther*. 2013; 138 (3):333–408. <https://doi.org/10.1016/j.pharmthera.2013.01.016> PMID: 23384594
12. Pushpakom S, Iorio F, Eyers P. A, Escott K. J, Hopper S, Wells A, et al. (2019). Drug repurposing: progress, challenges and recommendations. *Nat Rev Drug Discov*, 18(1), 41–58. <https://doi.org/10.1038/nrd.2018.168> PMID: 30310233
13. Schadt EE, Friend SH, Shaywitz DA. A network view of disease and compound screening. *Nat Rev Drug Discov*. 2009; 8 (4):286–95. <https://doi.org/10.1038/nrd2826> PMID: 19337271
14. Barabási A, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. *Nat Rev Genet*. 2011; 12:56–68. <https://doi.org/10.1038/nrg2918> PMID: 21164525
15. Yıldırım MA, Goh KI, Cusick ME, Barabasi AL, Vidal M. Drug—target network. *Nat Biotechnol*. 2007; 25:1119–26. <https://doi.org/10.1038/nbt1338> PMID: 17921997
16. Cheng F, Kovács IA, Barabási AL. Network-based prediction of drug combinations. *Nat Commun*. 2019; 10 (1):1–11. <https://doi.org/10.1038/s41467-018-07882-8> PMID: 30602773
17. Anighoro A, Bajorath J, Rastelli G. Polypharmacology: challenges and opportunities in drug discovery: miniperspective. *J Med Chem*. 2014; 57 (19):7874–87. <https://doi.org/10.1021/jm5006463> PMID: 24946140

18. Casas AI, Hassan AA, Larsen SJ, Gomez-Rangel V, Elbatreek M, Kleikers PW, et al. From single drug targets to synergistic network pharmacology in ischemic stroke. *Proc Natl Acad Sci*. 2019; 116 (14):7129–36. <https://doi.org/10.1073/pnas.1820799116> PMID: 30894481
19. Al-Lazikani B, Banerji U, Workman P. Combinatorial drug therapy for cancer in the post-genomic era. *Nat Biotechnol*. 2012; 30 (7):679. <https://doi.org/10.1038/nbt.2284> PMID: 22781697
20. Jaeger S, Igea A, Arroyo R, Alcalde V, Canovas B, Orozco M, et al. Quantification of pathway cross-talk reveals novel synergistic drug combinations for breast cancer. *Cancer Res*. 2017; 77 (2):459–69. <https://doi.org/10.1158/0008-5472.CAN-16-0097> PMID: 27879272
21. Matsunaga S, Kishi T, Iwata N. Combination therapy with cholinesterase inhibitors and memantine for Alzheimer's disease: a systematic review and meta-analysis. *Int J Neuropsychopharmacol*. 2015; 18 (5). <https://doi.org/10.1093/ijnp/pyu115>.
22. Guney E, Menche J, Vidal M, Barabasi AL. Network-based in silico drug efficacy screening. *Nat Commun*. 2016; 7:10331. <https://doi.org/10.1038/ncomms10331> PMID: 26831545
23. Misselbeck K, Parolo S, Lorenzini F, Savoca V, Leonardelli L, Bora P, et al. A network-based approach to identify deregulated pathways and drug effects in metabolic syndrome. *Nat Commun*. 2019; 10 (1):1–14. <https://doi.org/10.1038/s41467-018-07882-8> PMID: 30602773
24. Villeneuve DL, Angrish MM, Fortin MC, Katsiadaki I, Leonard M, Margiotta-Casaluci L, et al. Adverse outcome pathway networks II: network analytics. *Environ Toxicol Chem*. 2018; 37 (6):1734–48. <https://doi.org/10.1002/etc.4124> PMID: 29492998
25. Yeh SH, Yeh HY, Soo VW. A network flow approach to predict drug targets from microarray data, disease genes and interactome network-case study on prostate cancer. *Journal of clinical bioinformatics*. 2012; 2 (1):1. <https://doi.org/10.1186/2043-9113-2-1> PMID: 22239822
26. Piñero J, Ramírez-Anguita JM, Saúch-Pitarch J, Ronzano F, Centeno E, Sanz F, et al. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res*. 2020; 48 (D1):D845–55. <https://doi.org/10.1093/nar/gkz1021> PMID: 31680165
27. Darrah RJ, Jacono FJ, Joshi N, Mitchell AL, Sattar A, Campanaro CK, et al. AGTR2 absence or antagonism prevents cystic fibrosis pulmonary manifestations. *J Cyst Fibros*. 2019; 18 (1):127–34. <https://doi.org/10.1016/j.jcf.2018.05.013> PMID: 29937318
28. Kim MD, Baumlin N, Yoshida M, Polineni D, Salathe SF, David JK, et al. Losartan rescues inflammation-related mucociliary dysfunction in relevant models of cystic fibrosis. *Am J Respir Crit Care Med*. 2020; 201 (3):313–24. <https://doi.org/10.1164/rccm.201905-0990OC> PMID: 31613648
29. Vijayaraghavan S, Karakas C, Doostan I, Chen X, Bui T, Yi M, et al. CDK4/6 and autophagy inhibitors synergistically induce senescence in Rb positive cytoplasmic cyclin E negative cancers. *Nat Commun*. 2017; 8 (1):1–17. <https://doi.org/10.1038/s41467-016-0009-6> PMID: 28232747
30. Rocca A, Schirone A, Maltoni R, Bravaccini S, Ceconetto L, Farolfi A, et al. Progress with palbociclib in breast cancer: latest evidence and clinical considerations. *Therapeutic advances in medical oncology*. 2017; 9 (2):83–105. <https://doi.org/10.1177/1758834016677961> PMID: 28203301
31. Lee MS, Helms TL, Feng N, Gay J, Chang QE, Tian F, et al. (2016). Efficacy of the combination of MEK and CDK4/6 inhibitors in vitro and in vivo in KRAS mutant colorectal cancer models. *Oncotarget*, 7(26), 39595. <https://doi.org/10.18632/oncotarget.9153> PMID: 27167191
32. Adachi Y, Yanagimura N, Suzuki C, Ootani S, Tanimoto A, Nishiyama A, et al. Reduced doses of dabrafenib and trametinib combination therapy for BRAF V600E-mutant non-small cell lung cancer prevent rhabdomyolysis and maintain tumor shrinkage: a case report. *BMC Cancer*. 2020; 20 (1):1–4. <https://doi.org/10.1186/s12885-020-6626-9> PMID: 32093631
33. Tao Z, Le Blanc JM, Wang C, Zhan T, Zhuang H, Wang P, et al. Coadministration of Trametinib and Palbociclib Radiosensitizes KRAS-Mutant Non-Small Cell Lung Cancers In Vitro and In Vivo. *Clin Cancer Res*. 2016; 22 (1):122–33. <https://doi.org/10.1158/1078-0432.CCR-15-0589> PMID: 26728409
34. Simbulan-Rosenthal C. M, Dakshanamurthy S, Gaur A, Chen YS, Fang HB, Abdussamad M, et al. (2017). The repurposed anthelmintic mebendazole in combination with trametinib suppresses refractory NRASQ61K melanoma. *Oncotarget*, 8(8), 12576. <https://doi.org/10.18632/oncotarget.14990> PMID: 28157711
35. Hagberg A. A, Schult DA, and Swart PJ. (2008). Exploring network structure, dynamics, and function using NetworkX. *Proceedings of the 7th Python in Science Conference (SciPy2008)*, 11–15.
36. Staudt CL, Sazonovs A, Meyerhenke H. NetworKit: A tool suite for large-scale complex network analysis. *Network Science*. 2016; 4 (4):508–30.
37. Cowen L, Ideker T, Raphael BJ, Sharan R. Network propagation: a universal amplifier of genetic associations. *Nat Rev Genet*. 2017; 18 (9):551. <https://doi.org/10.1038/nrg.2017.38> PMID: 28607512
38. Grüning BA, Lampa S, Vaudel M, Blankenberg D. Software engineering for scientific big data analysis. *GigaScience*, 8(5), giz054. 2019. <https://doi.org/10.1093/gigascience/giz054> PMID: 31121028

39. Breit A, Ott S, Agibetov A, Samwald M. OpenBioLink: A resource and benchmarking framework for large-scale biomedical link prediction. *Bioinformatics*, btaa274. 2020. <https://doi.org/10.1093/bioinformatics/btaa274>.
40. Domingo-Fernández D, Mubeen S, Marin-Llao J, Hoyt C, Hofmann-Apitius M. PathMe: Merging and exploring mechanistic pathway knowledge. *BMC Bioinformatics*. 2019; 20:243. <https://doi.org/10.1186/s12859-019-2863-9> PMID: 31092193
41. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res*. 2017; 45 (D1):D353–61. <https://doi.org/10.1093/nar/gkw1092> PMID: 27899662
42. Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, et al. The reactome pathway knowledgebase. *Nucleic Acids Res*. 2020; 48 (D1):D498–503. <https://doi.org/10.1093/nar/gkz1031> PMID: 31691815
43. Slenter DN, Kutmon M, Hanspers K, Riutta A, Windsor J, Nunes N, et al. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res*. 2018; 46 (D1):D661–7. <https://doi.org/10.1093/nar/gkx1064> PMID: 29136241
44. Oughtred R, Stark C, Breitkreutz BJ, Rust J, Boucher L, Chang C, et al. The BioGRID interaction database: 2019 update. *Nucleic Acids Res*. 2019; 47 (D1):D529–41. <https://doi.org/10.1093/nar/gky1079> PMID: 30476227
45. Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, et al. The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res*. 2014; 42 (D1):D358–63. <https://doi.org/10.1093/nar/gkt1115> PMID: 24234451
46. Rodchenkov I, Babur O, Luna A, Aksoy BA, Wong JV, Fong D, et al. Pathway Commons 2019 Update: integration, analysis and exploration of pathway data. *Nucleic Acids Res*. 2020; 48 (D1):D489–97. <https://doi.org/10.1093/nar/gkz946> PMID: 31647099
47. Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, et al. (2010). DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic acids research*, 39(suppl_1), D1035–D1041. <https://doi.org/10.1093/nar/gkq1126>
48. Gysi DM, Valle ÍD, Zitnik M, Ameli A, Gan X, Varol O, et al. Network medicine framework for identifying drug repurposing opportunities for COVID-19. *arXiv*, 2004.07229. 2020. PMID: 32550253
49. Malas TB, Vlietstra WJ, Kudrin R, Starikov S, Charrouf M, Roos M, et al. Drug prioritization using the semantic properties of a knowledge graph. *Sci Rep*. 2019; 9 (1):1–10. <https://doi.org/10.1038/s41598-018-37186-2> PMID: 30626917
50. Berrar D, Flach P. Caveats and pitfalls of ROC analysis in clinical microarray research (and how to avoid them). *Brief Bioinform*. 2012; 13 (1):83–97. <https://doi.org/10.1093/bib/bbr008> PMID: 21422066
51. Hanhijärvi S, Garriga GC, Puolamäki K. Randomization techniques for graphs. In *Proceedings of the 2009 SIAM International Conference on Data Mining* (pp. 780–791). 2009. <https://doi.org/10.1137/1.9781611972795.67>.