HELMHOLTZ MUNICH

Modeling biomedical data with graph machine learning

Emy Yue Hu

July 2025

TECHNISCHE UNIVERSITÄT MÜNCHEN

TUM School of Life Sciences

Modeling biomedical data with graph machine learning

Emy Yue Hu

Vollständiger Abdruck der von der TUM School of Life Sciences der Technischen Universität München zur Erlangung des akademischen Grades einer

Doktorin der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz:

Prof. Dr. Dr. h.c. mult. Martin Hrabe de Angelis

Prüfende der Dissertation:

1. Prof. Dr. Maria Colomé-Tatché
2. Prof. Dr. Markus List

Die Dissertation wurde am 03.01.2025 bei der Technischen Universität München eingereicht und durch die TUM School of Life Sciences am 23.05.2025 angenommen.

Acknowledgments

In my PhD, I have learned so much scientifically and personally. I am truly grateful for this time, which was not always easy. Even more so, I want to express my gratitude to all who contributed significantly to this thesis.

Thank you, Annalisa Marsico and Nikola Müller, for giving me the opportunity to join your groups and learn so much from you. I enjoyed working in this interdisciplinary field, always expanding my horizons. You have taught me to be an independent scientist. Thank you for your guidance, supervision, advice, inspiration and ideas. Thank you, Maria Colomé Tatché, for collaboration and reviewing the thesis. Thank you, Martin Hrabe de Angelis and Markus List, for reviewing the thesis.

Thank you to my group, Janine, Christoph, Lambert, Ernesto, Marc, Svitlana, Ghalia, Samuele, Giulia, Patrick, Tobias, Andreina, Zahra, Hui. Discussions have been fun, and working together on a joint project was especially motivating.

Thank you, Lambert, for all the scientific discussions and strategic plannings of next steps and reality checks throughout the whole time! Thank you, Emilio, for your ears, your wisdom and encouragement. You gave me confidence and pushes to walk the paths I had to.

Thank you, Joachim Herz Foundation, for supporting my work with a fellowship that gave me the freedom to make independent plans, reaching for stars that I thought were out of my reach. Thank you, Causal Cell Dynamics, for making my trip to Mila possible where I found wonderful collaborators in Zhaocheng and Sophie. Especially, thank you, Sophie, for pushing me to ask questions when I was afraid to. It has been a great support to know such a great scientist had my back.

Thank you, ICB, with its director Fabian Theis, for providing an exciting environment to work and letting me meet some of the smartest, toughest, most hard working people around. I appreciate learning from you and growing with you: Anna, Johanna, Yannik, Akshaya, Karo, Lea, Barbara, Jonas, Natalie, Dominik, Alex, Juan!

Thank you to my family for your caring support in the brightest and darkest times of the PhD. Thank you to my partner in crime for your patience and your time, thanks for holding me accountable, especially in the last stretch.

Abstract

Biological function can rarely be attributed to an individual molecule; instead, it emerges at the system level through complex interactions involving proteins, DNA, RNA, metabolites, and other components. Similarly, disease can be understood as an emergent property derived from a deviation of regular interaction, influenced by external factors such as diet and stress. The advent of affordable and efficient high-throughput sequencing methods at an unprecedented scale has facilitated the investigation of disease states on many biological levels. Researchers can now study a disease based on multiple "omic" information layers, encompassing the complete set of gene expressions, methylations, genetic variants, and more. These layers allow for complex descriptions of biological processes. However, with the advancement of experimental techniques come challenges for computational tools to effectively utilize the wealth of available molecular patient data, recognize patterns, and draw conclusions, to advance the understanding of disease and develop innovative treatments.

To unravel the complex interaction patterns of heterogeneous biological entities, network biology has gained substantial attention. This interdisciplinary field combines life sciences, graph theory, and machine learning to provide deeper insights into biological systems. Biological networks naturally model biological entities as nodes and their relations as edges, such as protein-protein interaction networks describing physical interactions, and gene regulatory networks detailing causal relationships. Analysis and modeling techniques ranging from traditional network analysis methods and recent advances in graph-based machine and deep learning have already yielded substantial successes in biology to uncover disease-driving genes and detect gene modules.

The contributions of this thesis are three-fold: I extend and develop new frameworks to 1) integrate and 2) embed sequencing data across multiple omic layers, and to 3) predict missing interactions in complex heterogeneous biological networks. Specifically, this thesis introduces a multi-omic integration framework that leverages prior information, instead of analyzing each omic layer separately. The framework statistically selects features by modeling each response feature based on its multi-omic explanatory variables, thereby establishing a heterogeneous network. Applying the method to a patient cohort, key disease-modulating factors in major depressive disorder were found through node centrality measurements. Next, to delineate pathogenic mechanisms in complex COVID-19 infectious disease, this thesis presents a two-step network representation learning framework first to integrate and then embed all nodes into a latent space to efficiently explore the relationships between biological entities of genes, tissues, diseases, comorbidities, and genetically predisposed risks in COVID-19. Finally, this thesis contributes to closing gaps in our understanding of health and disease by developing BioKGC, an extension of a path representation learning algorithm, to predict missing links in biomedical knowledge graphs. Applications of BioKGC include predicting links to propose novel gene functions and repurposing of drugs as new therapeutic opportunities.

Zusammenfassung

Biologische Funktionen lassen sich nur selten auf ein einzelnes Molekül zurückführen; stattdessen entstehen die Eigenschaften auf der Systemebene in komplexen Interaktionen von Proteinen, DNA, RNA, Metaboliten und mehr. In ähnlicher Weise kann Krankheit als Eigenschaft des gesamten Systems verstanden werden, die sich aus einer Abweichung von der regulären Interaktion ergibt und durch externe Faktoren wie Ernährung und Stress beeinflusst wird. Mit dem Aufkommen kostengünstiger und effizienter Hochdurchsatz-Sequenzierungsmethoden in einem noch nie dagewesenen Ausmaß kann die Untersuchung eines Krankheitszustands nun auf vielen biologischen Ebenen erleichtert werden, indem mehrere "omic" Informationsebenen untersucht werden. Dabei beschreibt jede "omic"-Schicht den vollständigen Satz von Genexpressionen, Methylierungen, genetischen Varianten usw., was komplexe Beschreibungen von Prozessen ermöglicht. Mit den Fortschritten bei den experimentellen Techniken wachsen jedoch auch die Herausforderungen für Computer basierte Analyse Methoden, die die Fülle der verfügbaren molekularen Patientendaten nutzen und Muster erkennen können, um das Verständnis von Krankheiten zu verbessern und neue Behandlungen zu entwickeln.

Für die Analyse komplexer Interaktionsmuster heterogener biologischer Einheiten hat die Netzwerkbiologie als interdisziplinäres Gebiet, das auf Biowissenschaften, Graphentheorie und maschinellem Lernen basiert, besonders viel Aufmerksamkeit erlangt. Biologische Netzwerke modellieren biologische Einheiten als Knoten und ihre Beziehungen als Kanten, wie z. B. Protein-Protein-Interaktionsnetzwerke, die physikalische Interaktionen beschreiben, und genregulatorische Netzwerke, die kausale Beziehungen detaillieren. Die Netzwerkanalyse mit traditionellen Methoden und die neuesten Fortschritte im Bereich des maschinellen Graphenlernens haben in der Biologie bereits viel Erfolg gehabt, um krankheitsverursachende Gene aufzudecken oder Genmodule zu erkennen. Diese Doktorarbeit baut auf Prinzipien und Methoden aus diesem Bereich auf, um biologische Prozesse ganzheitlich mit allen omic-Schichten zu beschreiben und fehlende Interaktionen in komplexen heterogenen biologischen Netzwerken vorherzusagen.

Im Einzelnen wird in dieser Arbeit ein Rahmen für die Integration von Multi-Omics eingeführt, der alle Informationen für eine ganzheitliche Analyse nutzt, anstatt jede Omic-Schicht einzeln zu analysieren. Der Rahmen basiert auf der statistischen Auswahl von Merkmalen durch die Modellierung jedes Antwortmerkmals in Abhängigkeit von seinen multi-omic erklärenden Variablen, wodurch ein heterogenes Netzwerk entsteht. Anschließend wurden die wichtigsten krankheitsmodulierenden Faktoren bei major depressive disorder anhand von Messungen der Knotenzentralität ermittelt. Um die pathogenen Mechanismen der komplexen Infektionskrankheit COVID-19 zu beschreiben, wird in dieser Arbeit ein zweistufiger framework für das Lernen von Netzwerkrepräsentationen vorgestellt, um zunächst alle multi-omic Informationen zu integrieren und dann in einen latenten Raum einzubetten, um die Beziehungen zwischen biologischen Einheiten von Genen, Geweben, Krankheiten, Komorbiditäten und genetisch prädisponierten Risiken bei COVID-19 effizient zu untersuchen. Schließlich trägt diese Arbeit dazu bei, Lücken in unserem Verständnis von Gesundheit und Krankheit zu schließen, indem fehlende, aber wahre Verbindungen in biomedizinischen Wissensgraphen vorhergesagt werden, indem eine Methode zum Erlernen von Pfaddarstellungen angepasst und vorgeschlagen wird. Zu den Anwendungen gehören die Vorhersage von Verbindungen, um neue Genfunktionen vorzuschlagen und Medikamente als therapeutische Möglichkeiten neu zu positionieren.

Contents

Chapter 1

Introduction

Life, defined by the ability to breathe, eat, grow, move, reproduce, and sense, originated at least 3.7 billion years ago on our planet [Rosing, 1999]. While the earliest life forms were microbes, evolution has led to the emergence of many different organisms [Darwin, 1859]. As humans have always been curious about themselves and their surroundings, biology has developed as the empirical study of living organisms and their processes. It has revealed that life is organized into different hierarchical levels of biospheres, ecosystems, populations, individuals, organs, cells, molecules, and atoms [Urry et al., 2019]. Starting from atoms, new properties emerge with each higher organizational level [Urry et al., 2019]. The field of systems biology has been established, examining the regular arrangement of components that interact with one another, giving rise to the properties of an entire system [Kitano, 2002, Tavassoly et al., 2018].

At the individual level, health and disease are defined by the interplay of internal factors, such as molecules, cells, and organs, and influenced by external factors, such as diet and environmental cues [Mitchell, 2012]. In order to study health and disease, one possible research focus has been laid on the molecular level, studying the various interactions of biomolecules, such as deoxyribonucleic acid (DNA), ribonucleic acid (RNA), proteins, and metabolites, and their functions. For example, it has been shown that specific modifications to the accessibility of DNA are associated with expression changes of head and neck cancer-driving and growth factors [Kelley et al., 2017]. Ren et al. [2016] uncovered the association of metabolites with a loss of tumor suppressor genes in prostate cancer. As another example, for the neurodegenerative disorder Alzheimer's disease (AD), mutations in genes, such as amyloid protein precursor are thought to be responsible for the increase and aggregation of the amyloid β peptide in the brain [Lanoiselée et al., 2017], characteristic of the pathogeny of AD. Further, changes in the gene expression profiles are observed in patients, such as the upregulation of inflammation markers, e.g. tumor necrosis factor [Lima Giacobbo et al., 2019, Wang et al., 2015], while changes on the epigenetic level, such as through chromatin accessibility, help to explain how environmental factors change gene expression, including diet, stress, aging [Tecalco-Cruz et al., 2020]. In addition, oxygen and glucose hypometabolism in AD leads to changes in metabolites in the brain, such as reduced ATP availability for DNA repair processes or depletion of substrates for histone acetylation and subsequent transcriptional repression [Ardanaz et al., 2022, Bano et al., 2023, Bradshaw, 2021, Wu et al., 2023]. This shows that systems biology investigating health and disease is inherently interdisciplinary and complex [Burggren et al., 2017].

To make sense of this biological complexity and delineate the interactions, the more data we can use, the better. This has become feasible as better, faster, cheaper data collection techniques have become available and data is generated on an unprecedented scale. Sequencing the entire genome, epigenome, transcriptome, proteome, and metabolome has become feasible. Indeed, we can use all this "omic" information to comprehensively study a disease to facilitate understanding, identify biomarkers that are characteristic of diseases, and propose treatment options [Civelek and Lusis, 2014, Schadt, 2009, Yan et al., 2017]. However, at the same time, analyzing this wealth of data is challenging.

Network science, the study of complex networks, has gained popularity in modeling problems in diverse disciplines, such as social interactions, computer networks, language or politics [Ducruet and Beauguitte, 2014, Siew et al., 2019, Strogatz, 2001, Ward et al., 2011]. Rooted in discrete mathematics and building on methods from various disciplines, it describes the complex interactions of distinct elements (nodes, vertices or entities) mediated via their connections (edges or relations) [Barabási and Pósfai, 2016, Börner et al., 2007, Lewis and Lewis, 2009]. Network biology has naturally emerged as an essential topic area due to its ability to model the rich interactions between biological entities. Network analysis strategies have since been successfully applied to visualize [Shannon et al., 2003] and analyze [Barabasi and Oltvai, 2004, Lancichinetti et al., 2009] biological problems. They are based on the observation that biological molecules almost always act together rather than alone, such as in signaling pathways. Among the popular biological networks are gene regulatory networks, metabolic networks [Jeong et al., 2000, Ravasz et al., 2002], protein-protein interaction (PPI) networks [Rolland et al., 2014, Schwikowski et al., 2000], and drug-disease-gene networks [Huttlin et al., 2017].

1.1 Open challenges

Network biology is a fast-evolving field, due to the wide range of possibilities for application. With the improvement of theoretical foundations, methods and computational resources, the capabilities of network biology to model more and more complex interrelationships between entities has increased. However, still, many open challenges remain. Among many, the following details some important topics that will be addressed in this thesis:

(i) **Dynamic nature of biological systems:** Many available biological interaction networks are constructed by aggregating information across the scientific literature. These static networks oftentimes depict all possible interaction, e.g. between proteins that have been found over experimental screens. However, biological systems are dynamic and can vary depending on tissue or disease context. Thus, it is necessary to construct condition-specific networks that accurately depict the actual interactions within the system of study.

(ii) **Integrating diverse data types into networks:** The availability of high-throughput sequencing techniques allows the characterization of complex biological processes and diseases at a systems level. However, integrating diverse information, such as genomic, proteomic, metabolomic, and transcriptomic information (in general "omic" data), into a comprehensive network remains challenging. Main difficulties are the high dimensionality and heterogeneous data sources which vary in formats, scales and sequencing technique reliability.

(iii) **Efficiently exploring complex networks:** While network biology is powerful in elucidating the interactions of different molecular entities, another open question after network construction is how to efficiently explore and analyze these complex heterogeneous networks, intending to understand the interplay of factors and recurring network patterns and identify key players.

(iv) **Incomplete interactions:** Despite the effort of vast amount of experiments, biological interaction networks remain incomplete. These gaps in molecular interaction could lead to misleading conclusions. At the same time, closing gaps in our knowledge is essential, for instance, to explore gene function or disease treatment with drugs.

(v) **Disregard of specific relations:** Tools that operate on biological networks usually do not take the specific molecular interactions into account. Instead, they model relationships as homogeneous undirected edges. Taking the specific relation type into account offers more opportunities in modeling biological systems.

(vi) **Interpretability:** In the analysis of networks, interpretability is of high importance. Especially in biology, there is a need to understand why a prediction has been made to verify for biological plausibility and guide the way for new hypothesis generation. However, making sense of predictions to validate its credibility in biological networks is still a challenge.

1.2 Contributions of this thesis

With the rise of computational tools and, especially, with machine learning (ML) and deep learning (DL), we are able to tackle some of the challenges of complex and heterogeneous biological networks to gain information and deduce knowledge. In this thesis, I have developed and extended frameworks that (I) integrate various multi-omic patient/population data into heterogeneous networks, (II) efficiently explore these complex networks using node embeddings, and (III) close gaps in our biological understanding over link prediction:

(I) **Multi-omic network inference:** In the first contribution of establishing a framework for integrating multi-omic information, challenges (i) and (ii) are addressed, by creating a condition-specific network based on heterogeneous quantitative sequencing data of a Major Depressive Disorder (MDD) cohort. The framework relies on regularized regression approaches coupled with leveraging a prior network to deal with the issue of high data dimensionality. The model statistically selects explanatory features to predict a response variable, thereby establishing a heterogeneous network.

(II) **Multi-omic network embedding:** The second contribution recognizes the limitations of classical network analysis and addresses the challenges (i), (ii), and (iii) by efficiently exploring a complex multi-modal network derived from a population cohort to investigate the pathogenic mechanism of COVID-19 using network representation learning. The network nodes are projected to latent space to explore the relationships between tissue, gene, genetic predisposition, and phenotypes and elucidate the heterogeneous node context of COVID-19 genes.

(III) **Biomedical knowledge graph completion:** Last but not least, the third contribution addresses the challenges (iv), (v) and (vi) to close gaps in biological knowledge by predicting missing but true links in biomedical knowledge graphs by explicitly modeling the specific relation type between entities. Applications are the proposal of gene functions and treatments of diseases via drugs. An extension of a path-representation learning algorithm was established that allows to interpret findings, to verify predictions on their biological plausibility.

These contributions are part of scientific articles that have already been peer-reviewed and published, are currently under peer review, or are in preparation. Consequently, parts of this thesis are based on or closely aligned with the content of my following publications. Whenever the similarity between the text or images is nearly identical, the original publication is cited accordingly.:

- Ogris C., **Hu, Y.**, Arloth, J., & Müller, N. **Versatile knowledge guided network inference method for prioritizing key regulatory factors in multi-omics data**. *Scientific Reports, 2021*

- **Hu, Y.**, Rehawi, G., Moyon, L., Gerstner, N., Ogris, C., Bittner, F., Marsico, A., and Mueller, N.S. **Network embedding across multiple tissues and data modalities elucidates the context of host factors important for COVID-19 infection**. *Frontiers in Genetics, 2022*

- **Hu, Y.**, Oleshko, S., Firmani, S., Zhu, Z., Cheng, H., Ulmer, M., Arnold, M., Colomé-Tatché, M., Tang, J., Xhonneux, S., and Marsico, A.. **Path-based reasoning for biomedical knowledge graphs with BioPathNet.**, *biorxiv, 2024*

Other contributions from my doctoral research that are not directly used in this thesis are:

- Horlacher, M., Oleshko, S., **Hu, Y.**, Ghanbari, M., Elorduy Vergara, E., Müller, N.S., Ohler, U., Moyon, L., and Marsico, A. **Computational mapping of the human-SARS-CoV-2 protein-RNA interactome**. *NAR Genomics and Bioinformatics, 2023*

- Pardo, M., Offer, S., Hartner, E., ... **Hu, Y.**, ..., Czech, H., Kiendler-Scharr, A., Zimmermann, R., Rudich, Y. **Exposure to naphthalene and beta-pinene-derived secondary organic aerosol induced divergent changes in transcript levels of BEAS-2B cells**. *Environment International, 2022*

- Offer, S., Hartner, E., Di Bucchianico, S., ... **Hu, Y.**, ..., Czech, H., Kiendler-Scharr, A., Zimmermann, R., Rudich, Y. **Effect of atmospheric aging on soot particle toxicity in lung cell models at the air-liquid interface: Differential toxicological impacts of biogenic and anthropogenic secondary organic aerosols (SOAs)**. *Environmental Health Perspectives, 2022*

1.3 Thesis outline

This thesis is structured as follows: the remainder of the **introduction** details the relevant state of knowledge, as well as the essential scientific questions in the network biology research area. The background and motivation are given for each of the three contributions to the network biology field. Specifically, Section 1.4 details the need for a holistic view of biological systems, such as for the disease Major Depressive Disorder (MDD), and introduces method categories for multi-omic integration, focussing on selecting statistically relevant features. In Section 1.5, the complex disease pathology of Coronavirus disease 2019 (COVID-19) is presented, which calls for efficient analysis of integrated multi-omic data using network embeddings. Section 1.6 examines the need for an automated framework to predict links in biomedical networks, such as for functional annotation and drug repurposing. Further, it reviews the landscape of methods and finally calls for path representation learning for link prediction. Each introduction is wrapped up with a summary of the goals.

Next, in the **material and methods** chapter, the first four sections present the theoretical background. Section 2.1 introduces the basics of biological sequencing data that have become available for study through the latest technological advances, providing a comprehensive understanding of the foundation of our research. Sections 2.2, 2.3 and 2.4 further introduce the fundamentals of ML and DL, and graph ML, upon which the contributions of this thesis are built. Then, Section 2.5 discusses the methodological concept for integrating data from various sources into a multi-omic network based on statistical models leveraging prior information. They are followed by the specific material and methods used for the application of the method to a Major Depressive Disorder (MDD) patient cohort to identify key players in disease mediation. Section 2.6 presents the inference and embedding of a multi-modal network into latent space for efficient analysis of relationships between biological entities, such as genes and diseases. Then, it details the application to a pre-pandemic population cohort to contextualize COVID-19 genes in its multi-modal space, including clinical phenotypes, genomics, and transcriptomics across multiple tissues. Section 2.7 discusses adapting and using a path-based link prediction method for biomedical KGs to infer missing links. Subsequently, the application for the prediction of links between genes and pathways for functional annotation, and drugs and diseases for drug repurposing is presented.

In the following chapter, I present the most important **results** from each conducted study. First, the resulting multi-omic network derived from a quantitative patient MDD cohort is described. This is followed by a comparison to a two-omic integration technique. In addition, according to a centrality measure, the top 20 most important nodes are evaluated in their roles in MDD (Section 3.1). Second, for the contextualization of COVID-19 genes, the data from the population cohort with regard to genetic predisposition and comorbidities are explored at the beginning. Then, the latent space that results from the network inference and embedding framework is evaluated in its robustness, as well as in its meaningfulness based on tissue and disease-specific genes. Further, established host factors from various public experimental and patient data sources are investigated in their multi-modal context (Section 3.2). Third, the results of the application of the path representation learning tool for functional annotation and drug repurposing is presented, introducing novel insights. I show superior performance compared to embedding-based methods in the functional annotation task and on par and higher performance compared to the state-of-art tailored method for zero-shot drug repurposing, such as acute lymphoblastic leukemia, gastric

cancer and Alzheimer's disease. Examples of the interpretability tool to validate predictions in their biological plausibility are presented (Section 3.3).

In the **discussion** chapter, each contribution is summarized, and ideas for future methodological advancements and extensions are discussed for each of the three contributions (Sections 4.1, 4.2 and 4.3). This is followed by a cross-topic conclusion that discusses the potential and outlook of network biology in general (Section 4.4).

1.4 Network inference for multi-omic integration

In order to describe a complex biological system holistically, integrative approaches that analyze multi-omic data jointly are imperative (description of data modalities in Section 2.1). Multi-omic integration combines various sequencing techniques to study the interrelationship between biological entities giving rise to a disease. It is an active field of research with many efforts in method development, ranging from dimensionality reduction to feature selection (FS). Using the latter, such as in regularized regression (Section 2.2.3), the problem of high dimensionality can be tackled simultanously. Especially powerful are approaches that represent biological entities from multi-omic data as nodes in a interconnected network. This section is based on and partly identical to Ogris et al. [2021].

1.4.1 Disease complexity of Major Depressive Disorder (MDD)

Mendelian diseases have a simple genetic basis, primarily a single gene mutation. In contrast, complex diseases arise from the combined effects of multiple genetic and environmental factors [Spataro et al., 2017]. Multiple genes or proteins may contribute to the risk of disease development, with multiple genetic variations increasing or decreasing the risk. Environmental factors, such as lifestyle, diet, and adverse life events, may also contribute significantly to disease development. The interplay of these internal and external factors may further govern the development of a complex disease. MDD has a lifetime prevalence of 17–20 % and is an example of a complex disease [Kessler et al., 2005, Otte et al., 2016].

Genetic factors explain about 40% of MDD risk, with the remainder explained by individual-specific environmental exposures [Kendler et al., 2006]. Interestingly, genetic variations can modulate sensitivity and resilience to the long-term effects of adverse life events [Kendler, 2013] and modulate stress-induced gene expression, increasing MDD risk [Arloth et al., 2015b]

Therefore, research on complex diseases involves studying many different levels of information to identify the risk factors and their interplay, with the goal of prevention, detection, or treatment. With technological advances enabling the generation of vast amounts of sequence data from different biological levels, as introduced in Section 2.1, it is now possible to provide complementary views to understand a biological problem holistically, necessitating multi-omic analysis methods.

1.4.2 Multi-omic integration: Advantages and challenges

The simultaneous analysis of multiple types of biological data allows a comprehensive understanding of a complex system. Data modalities might include genomic, transcriptomic, methylomic, and phenotype information for each individual. This integration allows the investigation of how these biological entities interact and work together, leading to the system's functioning or malfunctioning. This way, robust results supported by different strands of evidence can be derived, complemented by findings unique to a specific data modality. Another advantage is reducing noise, which might arise from biological, experimental, and technical sources.

It is vital to analyze this convoluted mixture of different biological signals. However, the joint

analysis of multi-omic data presents many challenges. Therefore, one of the main goals of joint multi-omic analysis is to deconvolute this mixture and understand the underlying biological signals contributing to the overall state of the biological system.

1.4.3 Two-omic integration: Expression quantitative trait loci (eQTL)

One of the first steps towards an integrative analysis of more than one omic was the establishment of eQTL analysis, discovering genetic variations, such as SNPs, associated with gene expression variations. In other words, these genetic loci influence gene expression and can be regarded as regulatory variants, providing a deeper understanding of the molecular mechanisms underlying traits and diseases [Nica and Dermitzakis, 2013]. Many eQTL studies have been conducted, such as to investigate gene expression in the blood [Võsa et al., 2021] or tissues [Lonsdale et al., 2013].

$$expression = \alpha + \sum_{k} \beta_k \times covariate_k + \gamma \times genotype \tag{1.1}$$

In its simplest form, eQTL analysis involves using a standard additive linear model (Section 2.2.2) to explain expression levels based on genetic variation and covariates [Shabalin, 2012]. Multiple testing is accounted for by correcting the resulting p-value for the false discovery rate (FDR) using the Benjamini–Hochberg procedure. A distinction is also made between *cis* and *trans* eQTLs, which depends on the distance between the regulatory variant and the gene of interest. *Cis* regulatory variants are proximal to the gene, whereas *trans* variants are distal.

The same principle exists not only between SNPs and genes but also between methylation sites and genes. An expression quantitative trait methylation (eQTM) analysis seeks to uncover regulatory methylation sites [Kim et al., 2020].

1.4.4 Methods of multi-omic integration: Feature extraction vs. Feature selection

With efficient omic sequencing techniques, the number of studied features increases while the number of samples generally remains constant. When profiling the entire set of expressed genes and methylation sites, thousands of variables (p) might be measured from only a few hundred patients (n). This problem is called the curse of dimensionality, with $p \gg n$ increasing the problem of overfitting. However, the gain in information could outweigh the curse of dimensionality. In these cases, dimension reduction techniques are commonly employed [Mirza et al., 2019].

The most prominent methods for dimension reduction are feature extraction (FE) and feature selection (FS). FE aims to project from the high to a lower-dimensional space, such as via PCA [F.R.S., 1901] (Section 2.2.1.1). For multi-omic integration, popular methods include joint negative matrix factorization [Zhang et al., 2012] and multi-omics factor analysis (MOFA) [Argelaguet et al., 2018]. For example, MOFA decomposes the data matrices from different omics into a low-dimensional representation of samples and measurements. The inferred feature factors are linear combinations of the original features. This prevents the identification of the most relevant measured biological signal. Thus, one major downside of these latent representations is their low

interpretability.

In contrast, FS aims to reduce the high number of features into a smaller set of relevant variables, such as via LASSO [Tibshirani, 1996], elastic net [Zou and Hastie, 2005] (Section 2.2.3), and stability selection [Meinshausen and Buehlmann, 2009], where FS is an integral part of the model building process. In each model, the most relevant omic features are retained, e.g. to describe the expression of one gene. Meng et al. [2016] provide a good overview of dimension reduction algorithms and their applications in multi-omic integration.

Simultaneously, network analysis (Section 2.4) poses an important tool to understanding the interactions of biological entities that are involved in complex interactions essential for maintaining life in living systems [Tavassoly et al., 2018, Urry et al., 2019]. A method that operates on FS and then incorporates the selected features from the multi-omic spaces into a network would, thus, be valuable to capture complex interplay in disease and e.g. identify key players. During my Ph.D. studies, I contributed to a newly proposed integration method that finds relevant features in high multi-omic space and assembles them into networks to understand their complex interplay. In Section 2.5, I present the details of the method and its application to studying a cohort of patients with MDD.

1.5 Network inference and embedding to elucidate multi-modal node context for efficient analysis

As demonstrated in the previous section, multi-modal networks are essential to understanding the complex interactions in biological systems, such as in the psychiatric diseases of depression (Section 2.5). However, analyzing these networks with classical network analysis tools has become difficult as they become increasingly complex. Mining these networks to extract meaningful associations is important for understanding the underlying biological mechanisms. Here, graph representation approaches (Section 2.4.3) that generate node embeddings (i.e., representations of nodes as numerical vectors that reflect the network topology) have shown great promise for analyzing biomedical networks. The embedding space is optimized to learn the meaningful similarities and associations of the nodes in the network, aiming to preserve the proximity between a node and its neighborhood, and thus capturing its context. The interrogation of the embedding space is especially powerful for evaluating the relationship between nodes, characterizing the multi-modal context of any node in the network, such as existing diseases, phenotypes, genetic variations, and gene expression across a wide range of tissues. This section is based on and partly identical to Hu et al. [2022].

1.5.1 Disease complexity of COVID-19 caused by SARS-CoV2

The coronavirus strain called severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the main cause of COVID-19, a respiratory illness that caused the most devastating pandemic in recent history. After infection, most people experience symptoms such as shortness of breath, sore throat, cough, diarrhea, nausea, and fever. In severe cases, the disease progresses to a critical lung condition that leads to insufficient oxygen levels in the blood, named acute respiratory distress syndrome [Tay et al., 2020]. While respiratory system is primarily affected, numerous other tissues also exhibit a viral load [Demichev, 2021]. Consequently, studies have highlighted the impact of COVID-19 on various systems, such as the neurological, cardiovascular, hepatobiliary, and renal systems [Gupta et al., 2020]. Fatal cases of COVID-19 were recently associated with kidney and liver failure, further highlighting the crucial role of several chronic diseases in patient mortality [Elezkurtaj et al., 2021]. Other diseases associated with mortality from COVID-19 were alcohol and nicotine abuse, obesity, cerebrovascular disease, ischemic heart disease, chronic obstructive pulmonary disease (COPD) and arterial hypertension [Elezkurtaj et al., 2021]. Lung damage caused by a SARS-CoV-2 infection was associated with mortality rates but strongly predisposed by preexisting diseases (comorbidities).

Host genetics potentially contribute to COVID-19 pathogenesis with genetic loci found to be associated with an increased susceptibility to severe cases of COVID-19 [Group, 2020]. To explore the genetic component, the COVID-19 Host Genetics Initiative undertook a large-scale international effort, conducting GWASs that uncovered SNPs that were correlated with COVID-19 severity [The COVID-19 Host Genetics Initiative, 2020]. Collectively, these studies revealed that host antiviral defense mechanisms are affected by the genetic background, highlighting the diverse impact of the disease on various tissues and individuals. Understanding this complexity is crucial in the context of human diversity.

Alongside GWAS, functional experimental assays conducted in cell lines have provided insights into the molecular mechanisms involved in the response to SARS-CoV-2 infections. These studies investigated the host-virus interactions through ribonucleoprotein capture and immunoprecipitation [Gordon et al., 2020, Lee et al., 2021] to identify host factors that interact with viral proteins physically.

1.5.2 Need for holistic study using multi-omic data

Moreover, clustered regularly interspaced short palindromic repeats (CRISPR) studies have identified crucial SARS-CoV-2 infection host factors [Schneider et al., 2021, Wu et al., 2021]. Another valuable resource for understanding the viral response arises from quantifying differentially expressed genes, proteins, metabolites, and lipids in whole blood samples between COVID-19 cases and controls [Demichev, 2021, Di et al., 2020, D'Alessandro et al., 2020, Geyer et al., 2021, Messner et al., 2020, Overmyer et al., 2021, Shen et al., 2020, Wu et al., 2021]. While many studies have explored different aspects of SARS-CoV-2 infection, an integrated view is still lacking, especially the integration of the effects of infection across multiple tissues. Montaldo et al. [2021] recognized the significance of multi-omic studies in identifying pathogenic mechanisms underlying COVID-19 development based on a review of the domain literature.

Looking at the methods used for multi-omic data integration range include maximizing the correlation between layers of omic information [Singh et al., 2019] to multi-omic factor analysis performed in an unsupervised fashion [Argelaguet et al., 2018] to network inference [Ogris et al., 2021]. KiMONo, which belongs to the last group of methods, will be discussed in Section 2.5. In summary, KiMONo leverages prior information from existing biological networks to reduce high-dimensional input data and models every gene individually using a regularized linear model, such as a sparse-group LASSO. Then, the features selected from KiMONo's statistical models are aggregated into a network of different modalities by linking the explanatory variables to their modeled gene. However, the resulting multi-modal network is often highly complex, posing challenges for analysis using traditional network analysis tools such as degree and betweenness measures or module detection algorithms [Nelson et al., 2019].

1.5.3 Network embedding for efficient analysis of complex heterogeneous network

Graph representation learning methods have shown great promise in analyzing complex biomedical networks to mine meaningful associations from the network [Li et al., 2021, Nelson et al., 2019, Su et al., 2020], and have already been used for drug repurposing and predicting tissue-specific gene functions [Zitnik and Leskovec, 2017] and microRNA-disease associations [Li et al., 2017a]. These approaches optimize the geometry of the embedding space to capture meaningful similarities or associations between nodes within a given network, which can be used for downstream analysis using standard ML methods. Nelson et al. [2019] systematically compared classical network analysis strategies with graph representation approaches in network alignment, community detection, function prediction, network denoising, and pharmacogenomics. While they found that performance depended on the specific task and the metrics used for evaluation, they concluded that network embedding approaches have many advantages, including compu-

tational efficiency, especially for large and complex graphs, and robustness to inherent noise in biological networks.

Their greatest advantage is the continuous representation in vector space, allowing for vector operations and intuitive distance measures so that the optimized space can be used to infer relationships between nodes. For example, to investigate the multi-modal context of each node of interest, such as between genes and genetic risk scores or genes and tissues.

GeneWalk pioneered the prioritization of important connected nodes [Ietswaart et al., 2021], specifically in the context of genes and their biological gene ontology (GO) terms. The low-dimensional embedding space was learned from random walks over the heterogeneous network using the DeepWalk algorithm [Perozzi et al., 2014].

A list of differentially expressed genes is used as the input for network construction. A biological database is queried to obtain interactions between genes, and only nodes overlapping with the input gene list are retained. The network is then enriched with annotated GO terms, resulting in a heterogeneous network of genes and GO terms connected by unweighted edges. The unsupervised learning step applies the DeepWalk algorithm, using random walks to learn low-dimensional embeddings for nodes (Section 2.4.3.1). These embeddings capture the relationships between genes and their connected GO terms, enabling an informative characterization of the biological context.

Node embedding vectors are then used to calculate the cosine similarity score between nodes (Eq. 1.2) in the high-dimensional vector space, reflecting the similarity between genes and GO terms. Finally, the significance of the GO terms for genes is assessed by comparing similarity scores to a null distribution. Networks are randomized while preserving the node-degree distribution, and empirical p-values for GO terms are assigned based on comparing similarity scores with the null distribution.

$$\cos(\theta) = \frac{A \cdot B}{\|A\|_2 \, \|B\|_2} \tag{1.2}$$

The authors illustrated that this unsupervised representation learning algorithm's low-dimensional embedding enabled a characterization of each gene using it's annotated terms that were more informative in underlying specific biological context.

Recognizing both the need for an integrate multi-omic (or more generally multi-modal) view of the complex disease COVID-19 caused by the virus SARS-CoV2 and the potential of graph representation learning, we used ideas similar to GeneWalk to embed a heterogeneous condition-specific network which was derived from quantitative sequencing data of a population cohort to efficiently explore the relationships between entities in the context of COVID-19. I introduce the established two-step framework in Section 2.6 and detail the application to the case study of COVID-19.

1.6 Modeling network biology as relational graphs for link prediction

As outlined in Sections 2.5 and 2.6, multi-omic networks are powerful in depicting the interactions in biological systems. However, while the edges in these networks are only of one type (statistical association), in reality they may represent different relationships, such as physical interaction or of causal regulatory nature. This calls for the need to model the specific interactions between biological entities explicitly. Knowledge Graphs (KG) achieve this by encoding entities and their relationships as triplets, such as ("gene" – "involved in" – "cellular pathway", or "drug" – "treats" – "disease"). However, these KGs like our knowledge, despite scientific studies and consortia efforts, remains incomplete. Closing gaps in our knowledge, such as predicting the unknown function of genes or drug treatments, is of high importance. In the field of KG Completion which deals with the inference of missing yet true links, many techniques exist. While KG Embedding methods (Section 2.4.5) achieve good performances, one downside is the lack of interpretability of these models to guide verification and validation. In contrast, path representation learning (Section 2.4.6.3) has shown great promise in achieving higher performances and visualizing the most important paths of a prediction. This section is based on and partly identical to Hu et al. [2024].

1.6.1 Need for automated framework to close gaps in knowedge for functional annotation and drug repurposing

Edges in biological networks denote co-regulation or causal relationships (regulatory networks) [Karlebach and Shamir, 2008, Smith et al., 2007], physical interactions (in protein-protein interaction networks (PPI)) [Safari-Alighiarloo et al., 2014, Walhout and Vidal, 2001], or associations of genes with diseases (disease gene networks) [Goh et al., 2007, Piñero et al., 2020], among others. Even though a vast amount of high-throughput experiments have been conducted, many molecular interactions are not discovered, leaving gaps in our knowledge and biological networks incomplete. As wet lab experiments are bound to high time and energy investments, computational methods that perform link prediction (LP) based on the network topology have gained in importance to infer missing but potential associations between biological entities [Musawi et al., 2023]. The applications of LP in network biology reveals hidden connections between biological entities and facilitate the discovery e.g. of biomarkers and protein interactions [Abbas et al., 2021, Kumar et al., 2020, Wang et al., 2020].

Further of importance is the annotation of genes with thte functional role. It entails characterizing a gene regarding its biological function, molecular activities, and regulatory behaviors. The sheer number of genes and the expanding understanding of their roles in health and disease renders manual annotation via experimentation laborious and costly. In addition, diseases disrupt normal cellular behavior by affecting dozens of proteins and their interplay within the underlying biological network, changing the interactome [Huttlin et al., 2017]. In the best case, therapeutic drug treatment restores the intended molecular function and, thus, ameliorates the disease [Ruiz et al., 2021]. Consequently, automated annotation and drug prediction grounded in existing knowledge holds the potential for significant impact.

1.6.2 Homogeneous link prediction in biological graphs

One of simplest ways of LP between unconnected nodes is the usage of similarity metrics from traditional graph analysis, such as Personlized PageRank, Jaccard or Katz index [Jaccard, 1901, Katz, 1953, Page et al., 1999] in undirected graphs. These metrics have found successful application for the predicting associations between disease and gene [Iván and Grolmusz, 2011], and disease and drug [Lu et al., 2017], solely based on the topological similarity between nodes (Section 2.4.2.2). However, representation learning has shown higher expressiveness to capture the nuances and complexity of a node in a graph and thus have been discussed as a central method to network analysis [Nelson et al., 2019] (Section 2.4.3). Embeddings, low-dimensional vectors, are learned to represent each node in the network through shallow and deep non-linear transformation. The embeddings should summarize their graph position as well as the local graph neighborhood, so that geometric relations in the latent space correspond to graph relations. As an example, OhmNet predicts tissue-specific gene functions by learning embeddings via biased random walks (Section 2.4.3.1 on DeepWalk Perozzi et al. [2014]) for nodes in a multi-layer hierarchical network, where each layer represents the interaction of genes in different tissues [Zitnik and Leskovec, 2017]. Genes were then annotated with GO terms using a cosine similarity metric.

1.6.3 Heterogeneous link prediction in biological knowledge graphs

Early biological graphs do not capture the semantics of associations, e.g. in a PPI network unirelational associations do not capture the distinctions between interaction types such as inhibition and activation. Thus, recent efforts have turned to heterogeneous multi-relational networks to better represent biological complexities, which are known as knowledge graphs (KGs),(Section 2.4.4). A KG is represented in triplet format, detailing the connection between a head h entity via the relation r to the tail t entity as $\langle h, r, t \rangle$ with the entities and relations of different types. During KG Completion (KGC) the relation r is modeled specifically and missing links in the knowledge base are predicted like question answering given a query $p(?|h, r)$ to predict the t entity given the source node h and a specific relation r.

1.6.3.1 Node representation learning

For this task, KG embedding (KGE) methods (Section 2.4.5) have found great success, where the semantics of the triplets constrain the latent space. Node embeddings are randomly initialized and iteratively updated by optimizing a loss function to maximize the score for positive triplets and minimize that of negative triplets [Bordes et al., 2013]. In the embeddings space, TransE models the t as a linear transformation of h and r, while DistMult uses a dot product [Bordes et al., 2013, Yang et al., 2015]. Unlike the earlier methods, an encoder-decoder approach allows end-to-end training without prior graph feature extraction (e.g., over a random walk followed by a similarity metric). [Mohamed et al., 2021] benchmarked KGE methods on drug target interactions and tissue-specific protein functions, showing similar and higher predictive performances.

Graph deep neural architectures possess higher expressivity by aggregating node messages from neighbors to update the node's representation with representative methods such as Graph Convolutional Networks (GCNs) [Kipf and Welling, 2016] (Section 2.3.4.2), Graph Autoencoders

(GAEs) [Kipf and Welling, 2017] and GraphSAGE [Hamilton et al., 2017a]. Thus, as an improvement over KGEs, Relational Graph Convolutional Network (R-GCN) which is GCN operates on the multi-relational KGs by aggregating transformed feature vectors from the k-hop neighborhood separately for each relation type and subsequently uses the DistMult decoder to score labeled edges. Schlichtkrull et al. [2017] (Section 2.4.6.2).

1.6.3.2 State of the art method for drug repurposing

The models above perform LP on biomedical KGs compiled from various research efforts, with BioGRID, KEGG, and PathwayCommons as examples (Section 2.1.5). PathwayCommons unifies a range of different databases and details gene and protein interaction information with specific relations, such as "interacts-with," "in-complex-with," "controls-expression-of," and "controls-state-change-of" [Cerami et al., 2010, Demir et al., 2010, Rodchenkov et al., 2020]. Of particular note is PrimeKG [Chandak et al., 2023] (Section 2.1.5), a precision medicine KG designed explicitly to predict links between drugs and diseases. It incorporates over 120,000 nodes from 10 major biological levels of proteins, pathways, exposures, phenotypes, diseases, and drugs, detailing over four million relations. The fine-grained information about drugs and diseases, divided into `contraindications` and `indications`, makes PrimeKG an especially useful resource for drug repurposing. One state-of-the-art method based on node embeddings that predicts links between drugs and diseases is TxGNN. The authors attempted drug repurposing based on the protein interactome using a zero-shot prediction method, with little molecular characterization and no available information on treatment options. Huang et al. [2023] demonstrated the power of their model, TxGNN, using the PrimeKG dataset.

TxGNN uses a graph representation learning framework to embed the entities into a latent space optimized to reflect the geometry of the biomedical KG, focusing on diseases and their drug treatment candidates. The framework consists of these three important modules: 1) An encoder embeds the heterogeneous nodes of the graph, such as genes, pathways, diseases, and drugs, into a latent space constrained by the geometry of the KG. 2) A decoder with a metric learning module enriches the embeddings of zero-shot diseases with those from well-known diseases to make meaningful predictions for diseases with little molecular characterization and no known drug treatments. Given the embedding of a drug and a disease, TxGNN transforms them into a prediction about the relationship, such as indication and contraindication. 3) To learn the embeddings, training is conducted, first pre-training on all relationships in the KG and then fine-tuning the drug and disease edges.

R-GCN encoder The node embeddings are encoded following the R-GCN scheme (Section 2.4.6.2). First, the embeddings X_u for each node i are Xavier uniform initialized [Glorot and Bengio, 2010] during training. Next, for every layer in the message passing of the GNN (Section 2.4.6), messages are first passed, then aggregated, and used to update the node's embedding. The embeddings of the previous layer $h_u^{(k-1)}$ separately for each relation r are applied to a matrix $W_{r,M}^{(k)}$, thus obtaining the MESSAGE: $m_{r,u}^{(k)} = W_{r,M}^{(k)} h_u^{(k-1)}$.. Next, the function AGGREGATE average the incoming messages from each node v_u from the neighboring nodes $N_r(u)$: $m_{N(u)}^{(k)} = \frac{1}{|N_r(k)|} \sum_{v \in N_r(u)} m_{r,u}^{(k)}$. Finally, the node's embedding from the previous layer is updated by the UPDATE function of the MPNN $h_u^{(k)} = h_u^{(k-1)} + \sum_{r \in T_R} m_{N(u)}^{(k)}$. After k layers of propagation, the final embeddings for each node is obtained.

DistMult decoder The prediction likelihood of a given disease and drug embedding and the trained weight vector for each relation of interest is defined following the DistMult [Yang et al., 2015] (Section 2.4.5) interaction likelihood: $p_{u,v,r} = \frac{1}{1+\exp(-\sum(h_u w_r h_v))}$.

Metric learning model Each disease's signature embedding is generated based on its connectivity to other biological entities, such as proteins and exposures. The normalized dot product is calculated to obtain the similarities between two diseases, which can be understood as the amount of molecular entities shared. In the next step, the embeddings of zero-shot diseases are updated via auxiliary embedding, generated by recombining all similar diseases based on their degree. The more the disease has been characterized, the less it will rely on the auxiliary embedding.

1.6.3.3 Path representation learning

However, despite achieving good performances, disadvantages of LP methods using node embeddings are the low interpretability and the lack of capability to handle the inductive settings, whereby a graph with new nodes are supplied for inference. Thus, researchers have started to develop general and flexible LP representation learning frameworks based on path between two nodes. Zhang et al. [2023] mark a first application of this concept to prediction synthetic lethality (SL) gene partners. The authors introduce KR4SL, a GNN models with attention mechanism, leveraging a KG consisting of 24 relation types between 11 biological entities. Further, Neural Bellman-Ford Network (NBFNet) [Zhu et al., 2021] represents the paths of a node pair as the sum of paths, where each path is the product of edge representations (Section 2.4.6.3). A GNN with learned operators for efficient path formulation solutions is employed which is scalable to large graphs with low time complexity. NBFNet achieves better performance compared to node embedding methods and further offers the possibility of interpretability by visualizing important paths used for prediction. This in turn facilitates the verification of biological plausibility.

Seeing the advantages of path-representation learning, including higher performances and better interpretability, a framework using the principal and adapted to biomedical KG would, thus, be valuable for applications, such as functional annotation and drug repurposing. In Section 2.7, I present BioKGC, which is built upon NBFNet and adapted to biomedical needs for the prediction missing but true links. We benchmark the method thoroughly against node embedding based methods and especially TxGNN for drug repurposing.

Chapter 2

Material and Methods

In this chapter, I present the methods I (co-)developed for multi-omic integration, embedding and link prediction in biological knowledge graphs. For this aim, the first four sections present the theoretical background, from the basics of biological sequencing data that have become available for study through the latest technological advances (Section 2.1) over fundamentals of ML (Section 2.2) and DL (Section 2.3), to graph ML (Section 2.4). The contributions to network biology builds upon these background sections.

Section 2.5 presents the methodological framework for integrating data from multiple sources into a multi-omic network, utilizing statistical models that incorporate prior knowledge. This is followed by a detailed explanation of the materials and methods employed to apply this approach to an MDD patient cohort, aiming to identify key mediators of the disease.

Section 2.6 presents the inference and embedding of a multi-modal network into latent space, enabling efficient analysis of relationships between biological entities, such as genes and diseases. Subsequently, it details the application of this method to a pre-pandemic population cohort, contextualizing COVID-19 genes within a multi-modal space that includes clinical phenotypes, genomics, and transcriptomics across various tissues.

Section 2.7 discusses the adaptation and application of a path-based link prediction method for biomedical KGs to infer missing links. This method is then applied to predict links between genes and pathways for functional annotation, as well as between drugs and diseases for drug repurposing.

2.1 Biological Background

For the study of living organisms and their processes, biological data is meticulously collected and analyzed to synthesize hypotheses and knowledge. There are different ways of collecting data. Observations are made in animal and human behavioral studies to gain insights into instincts and learning. Experiments are conducted to understand the impact of interference or differential treatment in ecological conservation. Simulations are run to model behavior and predict future viral spread. In systems biology specifically, a wealth of molecular data is collected

to investigate health and diseases. These may include complementary information, such as an individual's genetic makeup, expressed genes, DNA modifications, and characteristics, such as age, body mass index (BMI), and sex. In addition, prior knowledge in the form of PPI networks is leveraged to inform the analysis and prediction.

Advances in modern sequencing techniques have enabled the large-scale, cost-efficient collection of such data. Multi-omic integration refers to combining data from multiple "omics" layers (i.e., different levels of biological organization) to study biological systems holistically [Santiago-Rodriguez and Hollister, 2021, Subramanian et al., 2020]. The task and challenge of data science in biology is to analyze this complex aggregate data to deduce patterns, test hypotheses, and predict behavior. While Section 2.2 introduces the methods used for the analysis, this section focuses on the fundamentals of each "omic" layer and the relevance to disease, followed by the experimental sequencing techniques. Further information about the fundamentals of biology can be found in Urry et al. [2019].

2.1.1 Genomics: DNA

The genome of an organism constitutes all of its genetic information. DNA has established itself as the information storage medium in most archaea, bacteria, plants, and animals due to its high stability. It is a double-stranded helix comprised of sequences of nucleotides. Each nucleotide consists of a sugar molecule, a phosphate group, and one of four bases: adenine (A), cytosine (C), guanine (G), and thymine (T). Nucleotides opposite one another in the helix are covalently bound via the 5' phosphate group of one and the 3' OH group of the other, forming the sugar-phosphate backbone of the DNA molecule. Among the bases, A pairs with T and C pairs with G, forming two and three hydrogen bonds, respectively, allowing the DNA to form a double helix. The human genome comprises roughly 6.3 gigabase pairs (Gbp) of DNA, but only around 1.5% of encode for 19,969 proteins. About twice as much DNA comprises non-coding RNA (ncRNA) genes, such as long noncoding RNAs (lncRNAs). There are also regulatory segments and introns, in addition to many sequences whose function remains to be discovered. Furthermore, significant segments of the genome contain gene duplications (6.61%), such as pseudogenes with limited function, as well as repeats (53.94%), once believed to be "junk DNA" [Nurk et al., 2022]. Many functions in gene regulation and disease have been discovered for both [Bustos et al., 2023, Jurka et al., 2007].

An individual's genome is inherited from both parents and is identical across all tissues. Germline mutations occur with a rate of 10^{-8} per nucleotide site per generation, resulting in around 100 de novo mutations in each newborn [Lynch, 2016]. It has been estimated that our genetics determine up to 80% of our susceptibility to and development of a disease, depending on its nature [Jackson et al., 2018, Klebanov, 2018, Sookoian and Pirola, 2017]. Thus, genetic predisposition plays a significant role in disease mediation. In addition, somatic mutations occur many orders more frequently than germline mutations during an individual's life, with one of many consequences being cancer [Lynch, 2016].

The methods for sequencing such extensive information started with Sanger sequencing in 1977 and now predominately involves next-generation sequencing (NGS) with its capabilities for massive parallel sequencing [Crossley et al., 2020, Quail et al., 2008, Sanger et al., 1977]. Goodwin et al. [2016] reviewed the different NGS approaches. Only mutations, called single nucleotide

polymorphisms (SNPs) and insertion and deletions of single or multiple base pairs that deviate from a reference genome, are stored for efficiency. One dominant data format is the variant call format, which describes the polymorphism's position in the reference genome and its reference and alternative alleles. Due to the nature of inheritance, various factors complicate working with DNA data. For example, corrections often must be applied at the cohort level to account for population structure or at the SNP level to account for linkage disequilibrium (LD), where neighboring alleles are correlated with one another [Reich et al., 2001, Sul et al., 2018].

2.1.2 Transcriptomics: RNA

The DNA segments that encode genes are synthesized into RNA through a process called transcription, which is performed by a multiunit enzyme called RNA polymerase. The entirety of the RNA products constitute the transcriptome. Unlike DNA, RNA is a single-stranded molecule in which ribose replaces deoxyribose and uracil (U) replaces the T nucleotide. Out of the roughly 20k protein-coding gene segments, more than four times the number of messenger RNA (mRNA) transcripts can be created due to post-transcriptional alternative exon splicing [Ryu et al., 2015].

A healthy cell state is characterized by coordinated gene expression, controlled via transcription factors, co-factors, and chromatin remodeling (see Section 2.1.3 for more information). Diseases arise due to a disturbance in cellular processes [Lee and Young, 2013]. In addition, gene expression constitutes the first response to internal and environmental cues. Extracellular signals activate signaling cascades, ultimately modulating transcription factors to mediate downstream gene regulation within the cell [Kabir et al., 2018]. Notably, ncRNAs, which are not translated into proteins, including microRNAs and lncRNAs, help to regulate gene expression, post-transcription modification, translation, protein degradation, and chromatin remodeling, among other processes [Klaff et al., 1996, Valencia-Sanchez et al., 2006, Wang et al., 2021b]. In summary, gene products play significant roles in health and diseases, where key disease genes are often differentially expressed.

Modern NGS technologies can be used to quantify gene expression. A complementary DNA library is created via reverse transcription of a sample's expressed RNAs before NGS amplification and sequencing. Not only can the presence and quantity of a transcript be determined, but post-transcriptional modifications and alternative exon splicing, among others, can also be profiled [Kukurba and Montgomery, 2015, Wang et al., 2009]. The data matrix can be used for analysis after quality control and normalization of expression values. Due to its high accessibility and informativeness about a sample's state due to both genetics and the environment, the transcriptome is one of the most studied information layers in biological systems.

2.1.3 Epigenomics: Methylation

Between the slow-changing DNA and the quick gene expression response, the epigenome governs the mid-term modification of inheritance that alters gene expression. Epigenetics describes modifications not to the DNA itself, but "on top" of it. They include the modifications of histone complexes around which the DNA is wrapped and the methylation of cytosine at a cytosine-guanine (CpG) site in the DNA. CpG sites are observed with less frequency than expected based

on the CG content of a human genome, commonly believed to be due to the high mutation rate. CpG islands are characterized by a CG content of more than 55% and are often found in gene promotors. The methylation of these CpG sites leads to gene expression repression that is not easily changed by cell division [Newell-Price et al., 2000, Takai and Jones, 2002].

Most epigenetic marks are obtained throughout the life of an organism, contributing to X-chromosome inactivation, response to environmental cues, and the differentiation of totipotent zygote cells into specific cell types through the activation and repression of specific genes [Feil and Fraga, 2012, Gendrel and Heard, 2014, Meissner, 2010]. However, some epigenetic changes can also be passed on to an offspring through transgenerational inheritance. In plants, these changes can prepare generations of offspring for dry climates. In humans, evidence suggests the inheritance of susceptibility to diseases such as cancer and obesity [Da Cruz et al., 2020, Fitz-James and Cavalli, 2022, Zheng et al., 2017]. Twin studies investigating the general effect of epigenetics on disease have been especially helpful, where individuals share the same genetic background but differ in their epigenetic profile, resulting in diseases such as cancer, autoimmune disorders, and depression [Javierre et al., 2010, Kaminsky et al., 2009, Mf, 2005, Penner-Goeke and Binder, 2019].

Standard sequencing techniques are chromatin immunoprecipitation and bisulfite sequencing. In the latter, double-stranded DNA is denaturalized into single strands and then treated with sodium bisulfite. While methylated cytosine is immune to bisulfite treatment, unmethylated cytosine is converted into uracil and is detected as thymine in subsequent polymerase chain reaction sequencing [Li and Tollefsbol, 2011].

2.1.4 Phenotypes: Sex, age, BMI, and comorbidities

Another layer of information that should be included in understanding health and disease is phenotypic data, such as age, sex, and BMI. The risk of diseases increases as tissue and organ functioning decreases with age, possibly due to diverse stressors, such as telomere erosion, reactive oxygen, DNA damage, and epigenetic stress [Childs et al., 2015]. Furthermore, some diseases have a sex prevalence. For example, there is a higher rate of neurodegenerative diseases in females or inherited via the sex chromosomes, such as in the case of X inactive specific transcript (*XIST*) [Loomes et al., 2017, Vegeto et al., 2020]. BMI also contributes to diseases such as cardiovascular or infectious diseases, including Coronavirus disease 2019 (COVID-19) [Kalligeros et al., 2020]. All of these phenotypic information are necessary to understand the disease and predict its further development. Other important information is the pre-existing diseases, called comorbidities, which can be associated with increased risk for developing a severe disease [Russell et al., 2023].

2.1.5 Biological interaction databases

When working with quantitative data derived from a patient cohort, it is highly beneficial to incorporate prior qualitative information, leveraging the accumulated knowledge from the scientific community. Relational databases that summarize the knowledge about the interaction of different biological entities are of high importance (more details on relational networks and

knowledge graphs in Section 2.4.4).

BioGRID is an example of a database that comprehensively curates protein-chemical, PPI, and other interactions [Oughtred et al., 2021, Stark et al., 2006]. The PPI comprises genetic and physical interactions derived from low- and high-throughput experiments. Indirect genetic interactions are captured via experimental evidence, such as dosage lethality, where the increased dosage of one gene causes the death of cells with a mutation in another gene. In contrast, direct interactions are measured in physical interaction experiments, such as two-hybrid or affinity capture-mass spectrometry, often involving bait and prey proteins. A detailed breakdown of all experimental systems can be found at wiki.thebiogrid.org. Similarly, PathwayCommons is another resource integrating sources detailing interactions between proteins and small molecules with a particular focus on their involvement in pathways and detailed interaction type, such as `in-complex-with`, `controls-expression-of`, or `controls-state-change-of` [Cerami et al., 2010, Demir et al., 2010, Rodchenkov et al., 2020].

A third example, PrimeKG, was built for precision medicine, going beyond protein interactions to incorporate further biological entities at ten major levels, such as anatomy, exposure, phenotype, and especially drugs and diseases, from 20 databases, including Disease ontology [Schriml et al., 2019], DisGeNET [Piñero et al., 2020], Orphanet [Weinreich et al., 2008], and DrugBank [Wishart et al., 2018]. The relationships detailed in PrimeKG include those between pathway-gene, exposure-disease, drug-gene, and especially drug-disease, broken down into `contraindication`, `indication`, and `off-label use`. Further details about the entities, relationships, and sources can be found in the PrimeKG article [Chandak et al., 2023].

2.1.6 Towards a holistic understanding of biological systems

In the nature and nurture model, genetics and environment together determine the development of diseases, which manifest at various biological levels [Day, 2000, Deedwania, 2004]. For a holistic understanding, multi-omic research focuses on the comprehensive set of molecules, such as measuring and analysing the genome in combination with transcriptome, epigenome and phenotypic information to shed light on the molecular underpinnings of diseases and the interplay of factors [Hasin et al., 2017]. In this endeavour, the knowledge accumulated by the scientific community in the form of biological networks can guide biomedical research in contextualizing and verifying results and generating hypotheses about disease mechanisms. This thesis used these databases as priors to inform the integration of multi-omic information during multi-omic network inference (Section 2.5 and 2.6), as well as direct input into ML frameworks to derive new links between entities (Section 2.7) closing gaps in our knowledge about biological systems.

2.2 Fundamentals of Machine Learning (ML)

ML and DL have become essential tools for analyzing complex data and making predictions in various scientific disciplines in the age of abundant data generated by modern measurement techniques and increasing computational power. In biology, the advances in these computational methods play a critical role in improving our understanding of the drivers behind biological processes, ultimately contributing to improvements in science and healthcare. This Section

will introduce key concepts in ML (Section 2.2) and DL (Section 2.3) that are applied and compared in this thesis. Then, it will present the fundamental concepts and notation of graph representation learning, reviewing classical graph analysis, graph embeddings, and graph neural networks (GNNs) used throughout this thesis (Section 2.4). The significance of methods for the modeling and analysis of biological data will be described in each section, especially in the section on graph ML (Section 2.4), focusing on network biology.

There are generally two popular approaches to studying any system. Mechanistic modeling represents causal processes and interactions within a system with usually simplified mathematical formulations [Ingalls, 2013]. This approach to modeling a system relies on a deep understanding of the underlying behavior of the system. In contrast, ML models are built on the principle that algorithms learn patterns through data, automatically adjusting internal parameters and improving performance. Thus, less domain expertise is required in ML. ML has been defined as a computer-based process whose performance improves with experience in a given task [Kitano, 2002, Mitchell, 1997]. However, the downside of ML compared to mechanistic models is their data-hungry nature and decreasing interpretability with increasing complexity. Baker et al. [2018] compares mechanistic modeling and ML comprehensively, especially in biological applications. ML is applied to many biological research questions due to advantages such as using structured and unstructured data and handling complex and large datasets [Greener et al., 2022].

ML can be broadly separated into supervised and unsupervised learning, which differ in their requirements and ingredients. In supervised learning, each data point has a label or value (more generally, a behavior), and the objective is to learn a model that can predict the behavior of the data. In unsupervised learning, such a behavior is missing, and the main objective is to uncover the underlying patterns, structures, or relationships in the data, such as to separate them into groups or clusters. The following section on unsupervised learning will touch upon representation learning that attempts to uncover underlying data patterns, with a straightforward linear model, principal component analysis (PCA), as an example. Then, supervised learning will be discussed in more depth, detailing linear to non-linear models and their optimization and evaluation. For more in-depth information on statistical learning and ML, please see Hastie et al. [2009], James et al. [2013], Murphy [2022]

2.2.1 Unsupervised learning

Unsupervised learning has several objectives, ranging from uncovering the underlying structure to anomaly detection to dimensionality reduction. It applies to many problems, such as investigating the semantic relationship between words or detecting fraud outliers. In biology, unsupervised learning can be applied for data exploratory purposes, patient stratification into treatment groups, or multi-omic integration to summarize data in fewer representative features. Standard algorithms for grouping are k-means or hierarchical clustering. Linear PCA or non-linear t-distributed stochastic neighbor embedding (tSNE) and uniform manifold approximation and projection (UMPA) are often used for dimensionality reduction.

2.2.1.1 PCA

PCA is a simple and widely used algorithm for dimensionality reduction that Pearson introduced in 1901 [F.R.S., 1901]. It aims to achieve lossy compression; that is, to minimize the memory needed to store the data with as little loss of precision as possible. Specifically, the aim is to replace many correlated variables with a smaller set of uncorrelated latent variables while explaining the bulk of the variance in the data. This reduction is especially important for biology since there are typically many highly correlated features. Reduced dimensionality can help with computational efficiency, visualization, and noise reduction.

The new uncorrelated latent variables are called principal components (PCs). The PCs are orthogonal, with the first explaining the greatest amount of variance, followed by the second, and so on. The number of PCs identified is typically limited using a cutoff for the percentage of variance in the data they collectively explain (e.g., 80%). The PCs are obtained through an eigendecomposition of the data covariance matrix, which is equivalent to the singular value decomposition (SVD) of the data matrix after normalization and centering. In SVD, a matrix is decomposed into singular vectors and singular values,

$$X = USV^T, \tag{2.1}$$

where U is a $n \times n$ orthogonal matrix, S is a $n \times p$ rectangular non-negative diagonal matrix, and V is a $p \times p$ orthogonal matrix. U and V represent the left and right singular vectors of X. The diagonal entries S_{ii} are unique until permutation and are called the singular values of X and are directly proportional to the standard deviations of the m singular vectors. The matrix V is called the loadings matrix or the principal directions/axes, with each column containing the linear combination of coefficients for each PC. The latter are given by

$$F = XV = USV^TV = US.$$

This matrix factorization successfully uncovers the structure underlying the data and simultaneously achieves a dimensionality reduction via lossy compression by selecting the top k PCs that cumulatively explain a given amount of the variance. There are many dimension-reduction techniques besides PCA.

2.2.2 Supervised learning: Linear models

Linear regression is a simple statistical learning approach from the supervised learning field. It is used to model and predict a quantitative response Y from a single predictor variable X, as in

$$Y = \beta_0 + \beta_1 X + \varepsilon, \tag{2.2}$$

or in the multi-variate setting when there is more than one predictor variable, as in

$$Y = \beta_0 + \sum_{j=i}^{p} \beta_j x_j + \varepsilon, \tag{2.3}$$

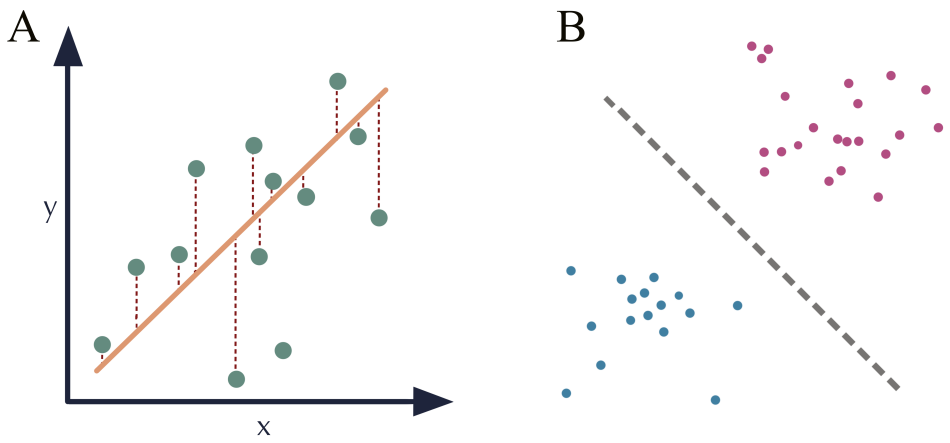


Figure 2.1: **Illustration of the two major tasks in ML:** A) linear regression representative for regression task and B) binary classification representative for label prediction.

where ε is the error term. The assumption is that there is approximately a linear relationship between the predictors and the response. Furthermore, other assumptions and requirements include the independence of errors with homoscedasticity and normality of residuals. The coefficients β are unknown and must be estimated such that the linear model fits the observed data. In other words, we want an intercept β_0 and slopes (β_1, β_n, ..., β_p) so that the resulting line is as close to the observed data points as possible.

Predicting Y for the i^{th} observation, denoted as $\hat{y}_i$, is achieved using the estimated coefficients in the linear regression model $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, where x_i represents the i^{th} value of the independent variable X. The estimates can be obtained by minimizing ε^2, also called the residual sum of squares (RSS).

$$RSS(\beta) = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=i}^{p}\beta_j(x_{ij}))^2 = e_1^2 + e_2^2 + \cdots + e_n^2 \tag{2.4}$$

as function of β.

2.2.3 Regularized linear models

Like model selection, regularization aims to identify the best model complexity for the data. In regularized linear models, such as ridge regression and least absolute shrinkage and selection operator (LASSO), parameters are selected via regularization techniques based on hyperparameters to manage complexity. The notion is that only a small subset of parameters (q out of p) are relevant. These methods fall under shrinkage techniques, wherein all p predictors are included, but coefficient estimates are controlled, reducing variance and preventing overfitting, thereby improving generalization. Importantly, regularization combats the curse of dimensionality, proving particularly valuable when the number of parameters is higher than the number of samples, $p \gg n$. This feature is essential in biology, specifically in multi-omic analysis, when there are tens or hundreds of samples but close to thousands or tens of thousands of measured variables, where it would be very advantageous to select the most essential parameters.

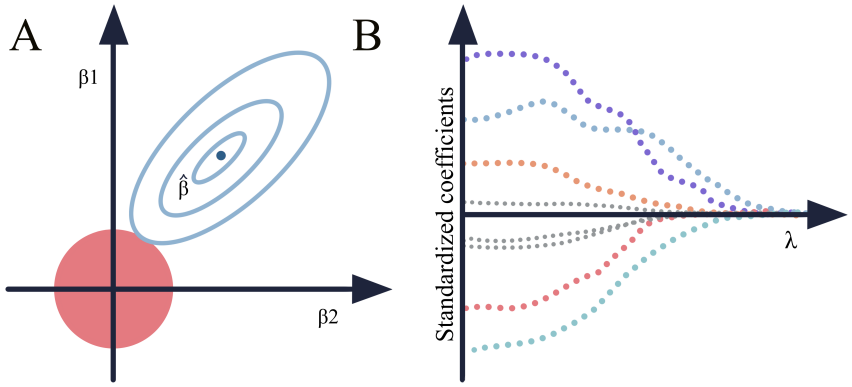


Figure 2.2: **Ridge regression:** ℓ_2 regularization with an illustration of the A) contours of the constraint function (red circle) and least squared error (blue ellipses) and B) decrease of the coefficients as the shrinkage penalty λ increases, but never reaching precisely zero. (Adapted from Hastie et al. [2009])

2.2.3.1 Ridge regression: ℓ_2 regularization

Ridge regression and LASSO are very similar. In the case of ridge regression [Hoerl and Kennard, 1970], the term that is minimized for least squares (2.4) is expanded by a penalty term acting on the coefficient estimates, giving us

$$RSS + \lambda \sum_{j=1}^{p} \beta_j^2, \tag{2.5}$$

where $\lambda \geq 0$ is a hyperparameter, called the shrinkage penalty, that controls the strength of the penalty. Ridge regression reduces to least squares estimates when $\lambda = 0$. However, higher penalties are applied when λ gets larger, and the coefficient estimates are pushed towards zero (Fig. 2.2). Therefore, the estimates shrink with higher λ. Ridge regression produces a new set of coefficient estimates $\hat{\beta}_\lambda$ for each value of λ. Then, the ideal set of estimates is selected based on cross-validation, which provides the best trade-off between variance and bias (see Section 2.2.6.1 for more information). Ridge regression is also referred to as ℓ_2 regularization since the penalty is applied via the ℓ_2 norm $||\beta||^2 = \sqrt{\sum_{j=i}^{p} \beta_j^2}$.

2.2.3.2 LASSO: l_1 regularization

Ridge regression only shrinks the coefficients towards zero while including all parameters in the final model and never setting them to precisely zero, which might be disadvantageous for model interpretability. This limitation is addressed by the LASSO [Tibshirani, 1996]. Here, the quantity that is being minimized is the following, where only the penalty term is replaced by the l_1 norm $||\beta||_1 = \sum |\beta_j|$:

$$RSS + \lambda \sum_{j=1}^{p} |\beta_j|. \tag{2.6}$$

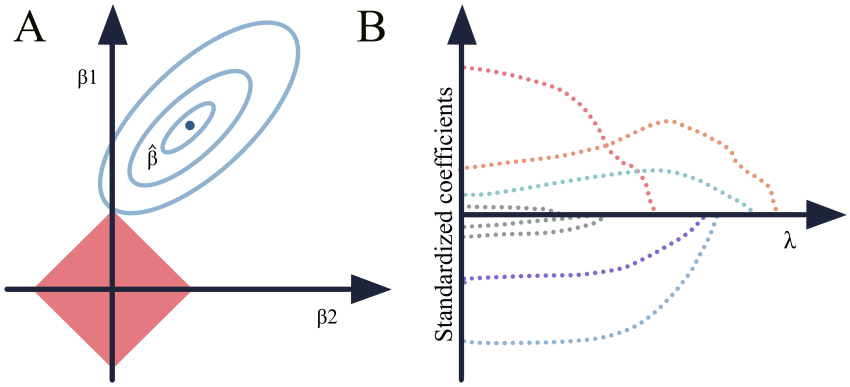


Figure 2.3: **LASSO regression:** ℓ_1 regularization with an illustration of the A) contours of the constraint function (red circle) and least squared error (blue ellipses) and B) decrease of the coefficients as the shrinkage penalty λ increases, reaching precisely zero with sufficiently large λ. (Adapted from Hastie et al. [2009])

Now, coefficients will be pushed towards zero and even set to precisely zero when λ is sufficiently large (Figure 2.3). LASSO will produce sparse models, which only consider a subset of variables. These models select one from multiple collinear features, which provides the advantage of lower complexity and, thus, better interpretability. There are many other regularization techniques besides the two most popular ones. For example, elastic net regression can be understood as a mixture of these two, combining shrinkage and sparsity-inducing properties [Zou and Hastie, 2005]. Group LASSO [Yuan and Lin, 2006] and sparse-group LASSO [Simon et al., 2013], which further apply a shrinkage (deactivation) on groups of features, are discussed in detail in Section 2.5.

2.2.4 Logistic regression

Logistic regression transforms linear models into non-linear models. Contrary to its name, the main task is not regression but classification (in the simplest form) into binary classes (Figure 2.4). The posterior probabilities of these classes are modeled over linear functions and ensured to sum to one, remaining within [0, 1], by passing the linear transformation of the data X with a weight matrix W (referred to as coefficient estimates β in linear regression) through a non-linear activation function σ.

$$F(x) = \sigma(XW + b), \tag{2.7}$$

with the sigmoid activation function given by $\sigma(z) = \frac{1}{1+e^{-z}}$, which converts the output of the linear regression equation into probability values between 0 and 1. The loss function that is to be minimized is

$$l(W) = \sum_{i=1}^{N} \{y_i \log(x_i w) - (1 - y_i) \log(1 - x_i w)\}. \tag{2.8}$$

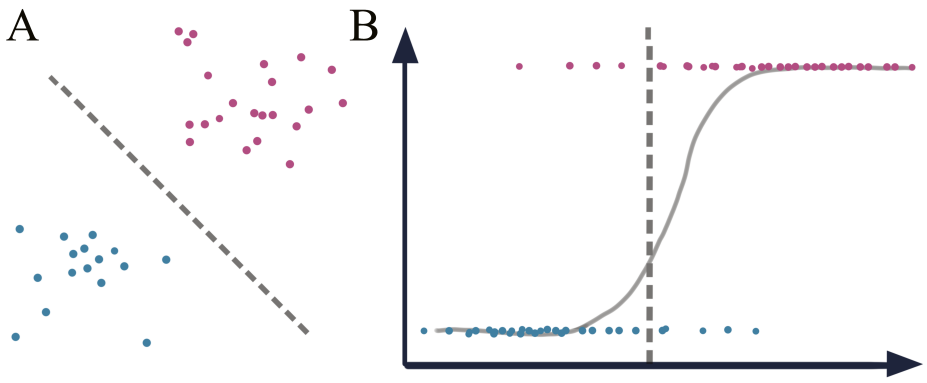


Figure 2.4: **Logistic regression:** A) Classification task of separating data points into two distinct classes, for instance, B) by using logistic regression - linear transformation followed by a non-linear activation function.

2.2.5 Model optimization

Optimization is the process of finding the best solutions from possible options. When optimizing ML models, we aim to minimize or maximize a specific objective or loss function. Generally, there are closed-form solutions and iterative numerical optimization strategies. While finding the best solution analytically over a closed-form solution is sometimes possible, iterative numerical approximations are more often feasible.

2.2.5.1 Maximum likelihood estimation

Besides minimizing the least squares, another popular method is maximum likelihood estimation, assuming the most reasonable values for the β are those with the highest probability for the observed sample. This optimization procedure is used when building probabilistic models, assuming a parametric for the distribution of $y|x$. The objective is to minimize the difference between the observed and assumed distribution as measured by the Kullbach-Leibler divergence. When solving the linear model, least squares and maximum likelihood estimation are equivalent when the error is distributed as $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, yielding

$$\mathcal{L}(\beta) = -\frac{N}{2}\log(2\pi) - N \ \log \ \sigma \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \beta(x_i))^2.$$

.

While there is also a closed-form solution to linear regression, given by $\beta = (X^T X)^{-1} X^T y$, the calculation can be expensive when inverting $X^T X$, especially in multivariate cases. Therefore, gradient descent, especially popular for DL, can be more favorable when the calculations are performed using parallelization across multiple processors.

2.2.5.2 Gradient-based optimization

2.2.5.2.1 First-order derivatives

Consider a function $f : \mathbb{R}^k \to \mathbb{R}$, $f(x) = y$, $k \geq 1$. If it exists, its derivative, denoted $f'(x)$ or $\frac{dy}{dx}(x) \in \mathbb{R}^k$, gives the infinitesimal slope of the function at the point x and is informative of how changes in the inputs affect the outputs y. When the inputs are vector-valued, $k \geq 1$, we can derive the partial derivatives $\frac{\partial f}{\partial x_i}(x)$, quantifying how f changes when only one of the variables, x_i, is increased. The gradient is a vector of all the partial derivatives, denoted as $\nabla_x f(x)$. The gradient of a function is zero, $f'(x) = 0$, when a stationary point is reached. Minima, maxima, and saddle points all satisfy this condition. In a *local minimum*, $f(x)$ is lower than all surrounding points, and gradient descent cannot further decrease the function by making infinitesimal steps. In a *global minimum*, $f(x)$ is not only the lowest value in the direct neighborhood but also its lowest global value. In optimization, we wish to minimize the loss function by iteratively moving the parameters according to the gradient, which is called gradient descent Cauchy and others [1847]. The new points proposed according to gradient descent are $x' = x - \varepsilon \nabla_x f(x)$, with ε as the learning rate.

2.2.5.2.2 Second-order derivatives

In addition to first-order derivatives (summarized as first-order optimization algorithms), second-order derivatives can also be applied, leveraging the function's *curvature* or *momentum* [Nocedal and Wright, 1999]. We can quickly determine whether a stationary point is a minimum, maximum, or saddle point using the second derivative, such as when examining the Hessian matrix. There may also be cases where the derivative decreases rapidly in one direction but slowly in another. Gradient descent algorithms are unaware of this distinction and will not preferentially explore the direction with the higher absolute derivative. Therefore, the step size might overshoot in one direction but be too conservative in another. An approach that considers this is Newton's method, which uses the Hessian matrix. Newton's method approximates the function near a specific point using a second-order Taylor series expansion, and this equation needs only to be applied once to converge to the minimum of a convex function. Newton's method (2.9) may need to be applied multiple times when dealing with non-convex functions that can be locally approximated as quadratic minima. However, it will converge to the critical point much faster than gradient descent.

$$x* = x^{(0)} - H(f)(x^{(0)})^{-1} \nabla_x f(x^{(0)}) \tag{2.9}$$

2.2.5.2.3 Example linear regression

Gradient ascent and descent are iterative optimization algorithms that find the extreme points of a function, such as the minima or maxima of a differentiable function. The starting point for the β coefficients is randomly initialized, the function's gradient is then calculated, and β is updated towards the direction with the steepest ascent or descent. The procedure is iterated until the convergence criteria are met.

No closed-form solution exists for logistic regression, and optimization can be performed over maximum-likelihood estimates like for linear regression. However, it is difficult to obtain an analytical solution due to the non-linear activation function. Thus, in most cases, gradient descent algorithms can be applied to derive the solution efficiently.

2.2.6 Model evaluation with performance metrics

2.2.6.1 Variance: Bias trade-off

When evaluating models, it is important to consider the trade-off between bias and variance (Figure 2.5A). Two competing properties of statistical learning methods govern the U-shape of mean squared error (MSE) curves. The expected MSE on the test x_0 can be decomposed into a bias, variance, and an error term:

$$E(y_0 - \beta x_0)^2 = Var(\beta x_0) + [Bias(\beta x_0)]^2 + Var(\varepsilon) \tag{2.10}$$

It would return the average test MSE if $\hat{y}$ were repeatedly estimated using many training sets, tested at x_0. While the irreducible error ε cannot be controlled, we want to find a model with low variance and bias. Variance refers to the amount the predictions that change if estimation is performed on different training datasets. If the variance is great, small changes in the training data yield large changes in the predictions. Bias refers to errors produced when a simpler model models a complex question.

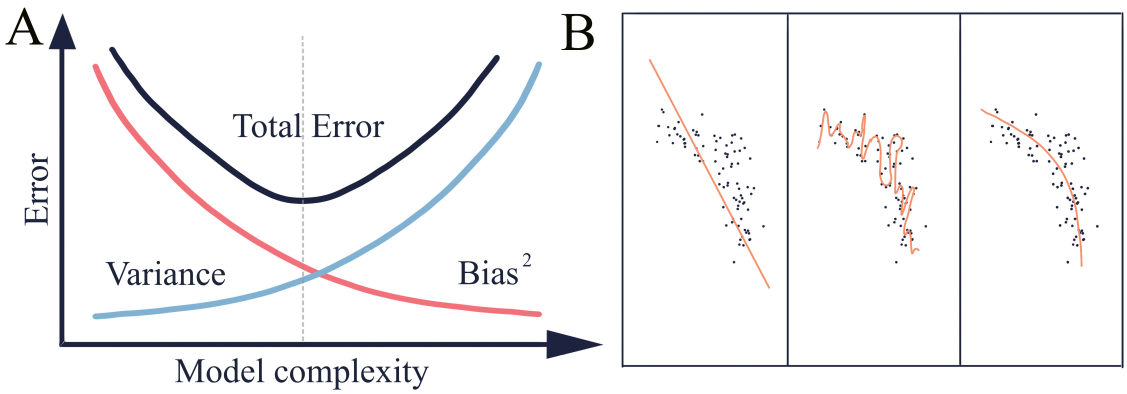


Figure 2.5: **Trade-off between variance and bias:** A) Relationship between the model complexity and total error made and B) underfitting (left), overfitting (middle) and a good fit (right). (Adapted from Hastie et al. [2009])

The complexity of a model, which is generally the number of predictor variables in linear regression or the number of parameters, is directly linked to these two terms. Overfitting occurs when the model is complex and accommodates every data point well. However, since many parameters are used, noise and random fluctuations are captured rather than the underlying relationship between variables (Figure 2.5B). This issue will lead to poor generalization since the model essentially memorized the training data, leading to poor predictions. In contrast, underfitting occurs when the model does not capture the underlying relationship adequately. The model is oversimplified, producing overly generalized predictions due to poor predictive powers.

The more complex a model, the higher the variance and the lower the bias, with the relative changes of these two parameters governing the MSE change. Their balance leads to the most favorable statistical learning model with the smallest MSE and the best generalization property to unseen data (Figure 2.5A).

2.2.6.2 Performance

The model's accuracy can be assessed using the MSE, which is RSS divided by the number of samples, and the variance explained R^2, quantifying its quality-of-fit in linear models. They are measures of the error or loss of the model. However, the commonly used metrics in classification are accuracy, area under the receiver operating characteristic curve (AUROC), and area under the precision-recall curve (AUPRC). While the loss can be investigated, accuracy measures the fraction of correctly predicted data points and is more meaningful in assessing performance. However, it depends heavily on the threshold when assigning the probabilities to the different classes. Thus, metrics that summarize the area under the curve are helpful when the threshold is set over the range of probabilities and the integral is calculated. They provide more information about the strengths and weaknesses of the model by assessing its accuracy over all thresholds. The AUROC provides the area under the curve for balanced classes, and the AUPRC is more meaningful for imbalanced classes. The baseline in AUPRC of a random classifier defaults to the class distribution.

2.3 Fundamentals of Deep Learning (DL)

As part of ML, DL was inspired by the functioning and structure of the human brain. The first artificial neural networks were modeled according to biological neural networks to simulate how humans process data. A few decades after its first invention, DL is now a highly relevant field of artificial intelligence (AI) to help process images (computer vision), understand and generate text (natural language processing), and make decisions (robotics and finance), and learns especially well from data such as images, text, audio, and graphs. The impact of DL in biology and biomedical research has also increased. Its applications include identifying tumors in medical images [Afshar et al., 2018], predicting protein binding to RNA [Horlacher et al., 2023], and classifying disease genes [Schulte-Sasse et al., 2021]. This section will review the fundamentals of neural networks, how they are optimized, and how domain knowledge is integrated. For more in-depth information, please see Goodfellow et al. [2016].

2.3.1 Neural networks

The *perceptron* was first introduced by Rosenblatt [1962] in the late 1950s to model the neurons of the brain. Like neurons in the brain, the perceptron receives signals from input sources and is activated when the signal XW is above the threshold $-b$:

$$F(x) = \sigma(XW + b) \tag{2.11}$$

It only deviates from logistic regression by replacing the sigmoid function with the Heaviside step function. Since the latter is not differentiable, gradient-based optimization methods cannot be used. However, the notion of a perceptron learning algorithm, randomly initiating weights and updating them according to the mistakes a model makes in its predictions, has had a massive impact. If the prediction is correct, no changes are applied; however, if the prediction is incorrect, the weights are updated in a direction to make the answers more likely. However, the persistent challenge of addressing many problems remains, including a strictly non-linear problem such as XOR, which involves learning to compute the exclusive OR of binary inputs within a function [Minsky and Papert, 1969]. To overcome this limitation, *multilayer perceptrons* (MLPs) were invented, a series of multiple perceptrons stacked on top of one another, and the Heaviside step function was replaced by a differentiable activation function. Common activation functions now include the rectified linear unit (ReLU) [Glorot et al., 2011] and its variations [Hendrycks and Gimpel, 2016, Maas et al., 2013]. The theorem states that an MLP, also called a feedforward neural network, is a universal approximator, meaning that it can closely approximate any continuous function with minimal error when sufficiently complex [Hornik et al., 1989]. Specifically, the hidden units z_l at each layer $l^{(i)}$ are obtained by applying a linear transformation to the hidden units from the previous layer $l^{(i-1)}$ and then passing the result through an activation function σ element by element:

$$z_l = f_l(z_{l-1}) = \sigma(b_l + W_l z_{l-1}) \tag{2.12}$$

These can be repeated recursively to create more complex functions of L layers with $f_l(z)$ as the function at layer l:

$$f(X, W) = f_L(f_{L-1}(\dots(f_1(x))\dots)). \tag{2.13}$$

2.3.2 Optimization

Neural networks are usually optimized using gradient-based approaches, using the concepts introduced in Section 2.2.5.2. The objective is to minimize the loss function to improve the model. Finding the global minimum is difficult due to the non-convex nature of complex loss functions due to their non-linearity. However, the global minimum is not of interest since the objective is to find good generalization properties (to discourage overfitting) at a low point. Instead, the training process can be stopped early by checking the validation error to find suitable parameterization. Optimizing DL models does not differ from training any other models discussed above, with the added complication that gradients are computed using back-propagation and its generalizations.

The cost function of an ML model is composed of the sum of the errors of all training examples, with the negative log-likelihood given by:

$$J(\theta) = \frac{1}{n} \sum_{i=1}^{n} L(x^i, y^{(i)}, \theta), \tag{2.14}$$

with the loss function $L(x, y, \theta) = -\log p(y|x; \theta)$. The derivatives must be computed across all training examples to derive the gradients, making one gradient descent step very expensive.

To overcome this, Rosenblatt [1962] used *stochastic gradient descent* (SGD), based on the insight that the gradient can be considered as an expectation over data points, which a small sample set, such as a *minibatch* of m samples, drawn uniformly from the training set can approximate. The model's weights can then be updated by:

$$\theta \leftarrow \theta - \varepsilon \frac{1}{m} \sum_{i=1}^{m} \nabla_\theta L(x^{(i)}, y^{(i)}, \theta), \tag{2.15}$$

$$\theta \leftarrow \theta - \varepsilon g \tag{2.16}$$

where $m \ll n$ and ε is the learning rate. There are m samples in one minibatch, and updating based on all minibatches represents one epoch. Another important factor is that the standard error of the mean of n samples is $\sigma/\sqrt{n}$, which yields less than linear returns with increasing samples. Therefore, many optimization algorithms would converge much faster by computing many approximate gradients than by slowly computing exact gradients.

Equation 2.15 is used to update the parameters of each layer, with their gradients derived by a recursive procedure called *back-propagation* [Rumelhart et al., 1986]. There are many extensions of SDG to improve the convergence, including using *momentum* [Polyak, 1964], various initialization strategies (e.g., normalized or sparse initialization) [He et al., 2015], second-order gradients (e.g., Newton's Method as briefly touched on in 2.2.5.2.2) [Martens and Grosse, 2015], and adaptive learning rates (e.g., AdaGrad [Duchi et al., 2011], RMSProp [Tieleman and Geoffrey Hinton, 2012], or Adam [Kingma and Ba, 2017]).

2.3.3 Strategies to overcome overfitting

Due to the high parameterization, neural networks generally face overfitting issues (see Section 2.2.6.1). There are a few strategies to counteract this. Weight decay applies l_2 regularization to network parameters [Loshchilov and Hutter, 2019], like in ridge regression (Section 2.2.3.1). In contrast, dropout randomly deactivates a portion of nodes in each layer before proceeding to the next layer, preventing the network from relying too heavily on specific pathways in the network [Srivastava et al., 2014]. Another way to reduce a model's variance is to introduce relational inductive bias [Battaglia et al., 2018].

2.3.4 Relational inductive bias

MLPs are the simplest types of neural networks that learn increasing abstract concepts along the layers to provide a final prediction output. While an MLP can approximate any function given a sufficient number of parameters [Hornik et al., 1989], it can be beneficial to use domain knowledge to introduce inductive bias into the network architecture, thus restraining the representational power to a domain of interest, gaining parametric efficiency, and improving generalization. The learning is biased towards structured representations by allowing the algorithm to prioritize one solution over the other [Battaglia et al., 2018]. This approach is crucial for images or the biological domain when learning from sequences or graphs. The relations are all-to-all in a fully connected network since each node in layer $l-1$ is connected to each node in layer l; thus, all nodes in the previous layer can interact to control the output (Eq. 2.11). The inductive bias is

higher for convolutional (CNN) and GNNs.

2.3.4.1 CNNs

CNNs were introduced to fields such as computer vision, audio processing, and time series [Lecun et al., 1998], and in biology for sequences [Budach and Marsico, 2018]. They are useful for identifying local patterns regardless of their positions in the grid-like input. CNNs are based on the operations of *convolution* and *pooling*. In the so-called *convolution* operation, much smaller *kernels* or *filters* are slid over an input (e.g., an image) to perform element-wise multiplication or summation and thus generate a *feature map* (Eq. 2.17). After applying several convolutions in parallel, the linear transformations are passed through a non-linear activation function, such as ReLU. Finally, a pooling function is applied, which summarizes the nearby outputs for a specific position, such as max pooling. CNNs are based on the principles of sparse interactions (locally in a filter) and parameter sharing (of learnable filter parameters), resulting in equivariant representations.

$$F^{conv}(x_{ij}) = \sigma(\sum_{(k=1)}^{K} x_{i+k-1} w_{jk} + b) \tag{2.17}$$

In CNNs, locality and translation *invariance* are imposed as important relational inductive biases. Locality is assumed, meaning that pixels close to one another in an image are related, while those further apart are not or less related. Translation invariance refers to the fact that features can be translated (e.g., rotated or shifted) but still be recognized by the model. While local invariance is achieved by applying convolutional filters across the image, translation invariance results from pooling operations.

2.3.4.2 Graph CNNs (GCNs)

GCNs are generalizations from CNNs to non-regular grids. As introduced by Kipf and Welling [2017], they operate on graphs and aggregate node features according to the neighbors of a node in the graph:

$$F^{graphconv}(X) = \sigma(D^{-\frac{1}{2}} \tilde{A} D^{-\frac{1}{2}} W X^T + b), \tag{2.18}$$

where D is the normalized graph laplacian and $\tilde{A}$ the adjacency matrix with added self-connections. While the convolutional operation is applied locally in CNNs, the node features are now aggregated from the neighbors of the graph, resulting in a relational inductive bias based on the graph structure. However, GCNs maintain the same translation invariance assumption as CNNs, namely that the detection of the pattern matters, not its exact location. The general GNN framework and message passing will be discussed in detail in Section 2.4.6.1.

Besides the mentioned architectures, other impactful models exist (but are not of great interest for this thesis), such as long-short term memory networks [Hochreiter and Schmidhuber, 1997] and transformers [Vaswani et al., 2017] for text processing, U-Nets [Ronneberger et al., 2015] for image segmentation, and generative diffusion models [Ho et al., 2020] for images, graphs and more.

2.4 Fundamentals of Graph ML

The origins of graph theory trace back to 1735 when the Swiss mathematician Leonhard Euler addressed the Königsberg Bridge Problem. This challenge involved finding a route through the city of Königsberg, crossing each of its seven bridges exactly once. Euler's solution introduced the use of graphs in mathematics, conceptualizing the city's geography as a network of nodes (landmasses) and edges (bridges), laying the foundations of graph theory [Alexanderson, 2006, Euler, 1741]. Research on graphs led to the development of many graph algorithms, especially in the twentieth century, such as depth-first search and breadth-first search [Cormen et al., 2022]. Another example is the Bellman-Ford algorithm for finding the shortest path from a source node to all other nodes [Bellman, 1958, Ford, 1956, Shimbel, 1954].

The versatility of (theoretical) graphs to represent and solve a wide range of problems resulted in the analysis of diverse (real world) networks [Barabási and Pósfai, 2016, Strogatz, 2001]. Examples include 1) social networks consisting of users and their diverse friendship, following, and collaboration relationships [Newman, 2001, Scott, 1992, Wasserman and Faust, 1994, Watts and Strogatz, 1998], 2) the internet as a technological network [Albert et al., 1999, Broder et al., 2000, Faloutsos et al., 1999], and 3) biological neural networks, food webs, and metabolic networks [Fell and Wagner, 2000, Jeong et al., 2000, Williams and Martinez, 2000]. Many algorithms have been developed to provide insights into the structure of various types of networks, with the overarching goal of understanding complex systems for improved decision-making or optimization.

This section will first discuss graph theory (2.4.1) and its classical graph analysis strategies (2.4.2), and then move on to node representation learning (2.4.3), methods for link prediction using node embeddings (2.4.4), and path representation learning (2.4.6). See Barabási and Pósfai [2016], Hamilton [2020] for a more detailed description. Each subsection will also investigate how graphs are used to analyze biological data, from discovering key players to elucidating the context of nodes to link prediction in biomedical KGs.

2.4.1 Graph theory

A graph $G = (V, E)$ is defined by its sets of nodes V and edges E, where an edge $e = (u, v) \in E$ leads from node u to v with $u, v \in V$. Common ways to represent graphs are *edge lists* or *adjacency matrices* $A \in \mathbb{R}^{N \times N}$ with $N = \#V$:

$$A_{u,v} = \begin{cases} 1, & \text{if } (u, v) \in E \\ 0, & \text{otherwise} \end{cases} \tag{2.19}$$

If the edges are *undirected* (as opposed to *directed*) (Figure 2.6), the direction of an edge is not important, $(u, v) \leftrightarrow (v, u)$, and the adjacency matrix used to represent the graph is symmetric. Furthermore, when the entries in the adjacency matrix entries are real-world or normalized values instead of binary 0 or 1, the graph is regarded as a weighted network. These *edge features* may reflect the strength of interactions between two nodes. In addition, different *edge types* $f : E \to R$ might occur, where R is the set of types that exist. Graphs can also be classified into *heterogeneous* networks (as opposed to *homogeneous*) when multiple node types are present.

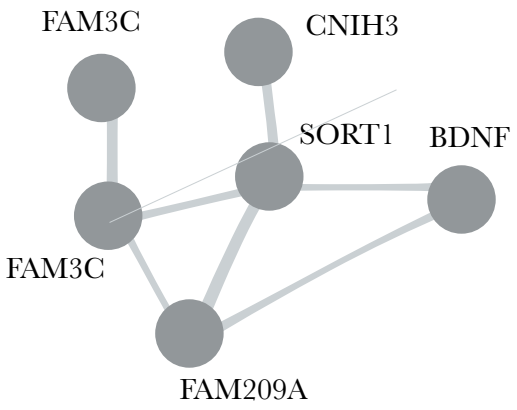


Figure 2.6: **Example of a biological network:** PPI derived from yeast-two hybrid screens with undirected and homogeneous edges.

Furthermore, graphs can have *node features* of scalar values or vectors associated with nodes, reflecting some node properties. Another important concept in graph theory is the *path* that is a distinct sequence of nodes joined by a distinct sequence of edges. They are essential for the understanding of the connectedness of the graph. Relatedly, *communities* are local densely connected subgraphs in a network with all members reachable through other members of the same community. As an extreme, a *clique* is a set of nodes such that every two nodes in the clique are connected by a unique edge, forming a complete graph of interconnectedness. While *strong communities* satisfy the constraint that every node has more links within the community than towards the rest of the graph, *weak communities* only need to possess in total more internal than external links, allowing for individual nodes to violate the constraint.

Biological networks are of various types. As an example, PPI networks (mentioned in Section 2.1.5) that detail the physical interaction between proteins derived from experiments are undirected with the symmetrical relationship of "interaction". Other networks capture nodes of various types (such as `proteins`, `pathways`, `diseases`, or `drugs`) as well as edges of various types (such as `phosphorylation`, `cellular function of`, `disease protein` or `indication`). Here, edges usually satisfy the constraint that certain edges only occur between certain node types, as is the case when proteins are connected by PPIs, where functional annotation occurs between proteins and pathways or treatments between drugs and diseases. In this thesis, Section 2.5 and 2.6 deal with undirected heterogeneous networks in terms of nodes, and Section 2.7 with directed heterogeneous networks in terms of nodes and edges. It is assumed that biological networks—like many real-world networks—typically follow the scale-free property with hubs and a skewed degree distribution with a heavy tail. They are also assumed to follow the small-world assumption, where, despite the seeming complexity of a large network, any two nodes are actually not far apart from one another in the network [Strogatz, 2001].

2.4.2 Classical network analysis

Common graph tasks include classifying nodes, classifying graph levels, predicting relations between nodes, and detecting clusters [Hamilton, 2020]. Before modern DL graph approaches,

various classical graph learning strategies were developed and are still in use today [Ashtiani et al., 2018]. This section will review the classical approaches before moving on to modern methods relying on representation learning.

2.4.2.1 Node classification

One fundamental task in network analysis is classifying nodes regarding their importance or property. In social networks, we can identify nodes that are important or influential to the entire graph. For example, the position of the Medici family in the network of fifteenth-century Florentine marriages revealed them to be powerful [Freeman, 1978, Padgett and Ansell, 1993]. **In computational biology**, we might be interested in identifying essential key players in disease, health, or protein interactions [Ashtiani et al., 2018] or classifying genes by their disease involvement, such as in cancer [Schulte-Sasse et al., 2021].

There are useful statistics characterizing the nodes in a graph to determine a node's importance. *Node-level statistics* include *degree* and *centrality*. Degree examines the number of edges connecting a specific node, denoted as u, to other nodes in the set V:

$$d_u = \sum_{v \in V} A[u, v] \tag{2.20}$$

The degree of a node thus quantifies its number of neighbors. In a directed graph, the distinction is made between incoming and outgoing edges. It is calculated by summing over the respective rows or columns of A.

While the degree is informative about a node's number of neighbors, centrality provides a different view of a node's role in the graph. One popular measure of the overall connectedness is the *eigenvector centrality*, which ranks nodes according to the likelihood of being visited on an infinitely long random walk. Another is *betweenness*, which focuses further on occupying pivotal positions in the connectivity of networks. It is calculated as

$$b_u = \sum_{u \neq v \neq w} \frac{\sigma_{uw(v)}}{\sigma_{uw}} \tag{2.21}$$

where σ_{uw} is the total number of shortest paths between nodes u and w and $\sigma_{uw(v)}$ is the number of shortest paths between u and w that pass through node v. Therefore, betweenness measures how often a node lies on the shortest path between other nodes, exhibiting a key position in the flow of information in the network. More information can be found in Newman [2018].

In the biological setting, other measures that have been evaluated by Koschützki and Schreiber [2008] in their role to identify key regulators in transcriptional control networks of *Escherichia coli* are closeness centrality (reciprocal sum of length of shortest paths), Katz centrality (eigenvector centrality taking into account non-direct neighbors) [Katz, 1953], PageRank (importance based on the number and importance of incoming edges) [Page et al., 1999] and motif-based centralities [Koschützki et al., 2007]. The authors showed that more than 50 % of the important global regulators were retrieved in the 2% of predicted ranked genes. Interestingly, while all metrics measure centrality, each covers different information, with correlations among them sometimes dropping as low as 0.14.

2.4.2.2 Link prediction

Another important task in network analysis is predicting links between two nodes that are not currently present, but could be true or relevant. In social networks, link prediction can uncover potential friendships between users [Liben-Nowell and Kleinberg, 2003]. Inferring missing links in networks built on co-authorship and institutional connections might identify future collaborations [Wang and Sukthankar, 2014]. **In biology**, link prediction is performed, e.g., to infer interactions between proteins for regulatory prediction or between drugs and diseases for the uncovering of new therapeutic opportunities [Musawi et al., 2023]. The wide range of applications makes predicting the likelihood of a connection based on the structure and pattern of the existing links important.

Several methods have traditionally been used to determine the existence of an edge between two nodes, including computing similarities based on the overlap of the local neighborhood (Sorensen and Jaccard index [Jaccard, 1901, Sørensen, 1948] and global overlap measures (Katz index and improvement thereof [Katz, 1953, Lü and Zhou, 2011]). While the former considers the overlap of common neighbors, the latter takes the number of paths of any length (compared to the expected number) between a pair of nodes. A third category of methods relies on random walks, such as the extension of PageRank named the personalized page rank (PPR) [Page et al., 1999]. It provides the probability for a random walk starting at source node u and ending at target node v, which, intuitively, measures the bidirectional importance between two nodes. By summing the PPR scores obtained from node u to v and vice versa, the similarity metric can reflect how likely it is to move between them and, thus, how likely a link is to exist.

In biology, the PPR has found application in the analysis of metabolic and protein networks [Iván and Grolmusz, 2011]. The PR algorithm was personalized to 13 selected proteins with high expression levels in plasma, and the authors recovered highly related proteins from a total of over 27k proteins. Other studies have studied the benefit of using similarity indices, such as common neighbors, Jaccard or Katz index, for drug repositioning [Lu et al., 2017].

2.4.2.3 Clustering on graphs

Another way to extract information from a network is a fundamental technique called clustering, which aims to uncover patterns and structures with a higher local density than a randomly wired network [Barabási and Pósfai, 2016]. Characterizing a network through its modules, which are sets of highly interconnected nodes, could reveal joint properties, mechanics, or functions. In social networks, clustering revealed groups of people that interact more frequently with each other than with others outside of the group [Handcock et al., 2007, Mishra et al., 2007]. **In the analysis of biological networks**, modules uncovered in protein networks corresponded to protein complexes that physically interact or are active in the same signaling pathway [Alcalá-Corona et al., 2021, Girvan and Newman, 2002, Hintze and Adami, 2008].

We will examine some fundamentals of spectral graph theory to identify clusters of nodes in a graph. The *graph Laplacian* transforms the adjacency matrix A and holds useful mathematical properties:

$$L = D - A, \tag{2.22}$$

where D is the degree matrix. For example, the number of eigenvectors with an eigenvalue of 0 corresponds to the number of connected components in the graph, which is a set of nodes linked through edges and paths. The eigenvector indicates which nodes are in the connected component. The symmetric normalized Laplacian used by GCNs already appeared in Section 2.3.4.2.

$$L_{normalized} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}} \tag{2.23}$$

Clustering based on the connected component is trivial. According to the Rayleigh-Ritz theorem, we can use the second-smallest eigenvector of L to cluster within a connected component, which provides a continuous approximation to discrete optimal cluster assignment, minimizing cross-cluster edges. Ravasz and Barabási [2003] and Newman and Girvan [2004] proposed methods for community detection that are based on hierarchical clustering with a similarity matrix as a foundation. They use agglomerative and divisive algorithms, respectively, by either merging nodes into the same community when a high similarity is observed or removing nodes in low similarity. Other algorithms, based on random walks include Louvain [Blondel et al., 2008], Infomap [Rosvall and Bergstrom, 2011] alongside with Label propagation [Raghavan et al., 2007] and Walktrap [Pons and Latapy, 2005].

In biology, it has been postulated in *networks of processes* theory that modularity is necessary to attain the sophisticated degree of organization seen in living systems [Clarke and Mittenthal, 1992]. Apart from finding protein modules involved in the same signaling pathway, the identification of diseases subnetworks or cancer driver discovery in multi-network communities has also been of focus [Barabási et al., 2011, Cantini et al., 2015, Ghiassian et al., 2015, Menche et al., 2015].

2.4.3 Node representation learning

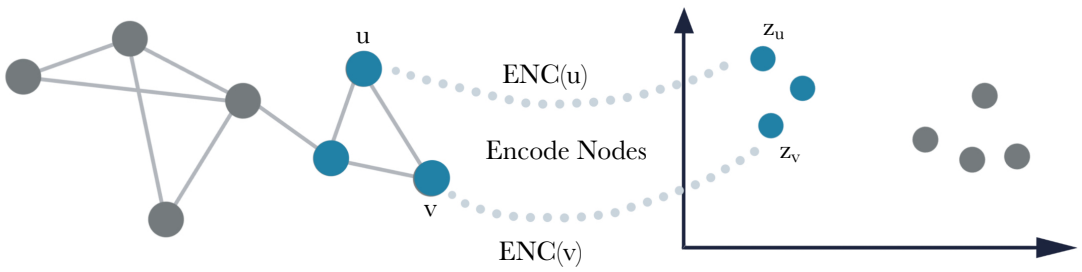


Figure 2.7: **Projecting nodes of a network into an embedding space** which is optimized so that the distances in the latent space is reflective of the node's position in the network (Adapted from Hamilton [2020])

In addition to traditional network analysis methods summarizing network properties into hand-engineered statistics, learned representations can adapt by learning from data, providing greater flexibility. Graph embeddings have gained importance as a technique to project nodes onto a latent space, enabling more efficient analysis of complex networks. The primary objective of learning node embeddings is to summarize information about a node's position and the structure of its local neighborhood into a low-dimensional vector (Figure 2.7). The aim is to capture the

graph's relationships as geometric relations within this latent space [Hoff et al., 2002]. Then, the resulting embeddings can be used for subsequent tasks, such as investigating node similarity, cluster detection, or prediction of links. In the application of network embedding to **biological problems**, network embedding has been shown to be advantageous for a range of problems from the prediction of functions, network denoising, and pharmacogenomics [Nelson et al., 2019].

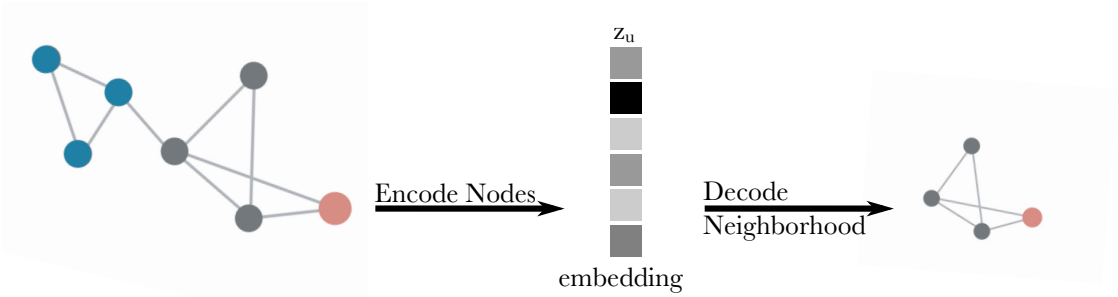


Figure 2.8: **Encoder and Decoder approach in network representation learning** where the encoder learns a representation of a node u given by a low-dimensional vector z_u and the decoder uses the embedding to reconstruct the local neighborhood. (Adapted from Hamilton [2020])

A useful way of thinking about learning node embeddings is applying the encoder-decoder concept. The encoder is responsible for projecting a node $v \in V$ into d dimensional latent space, obtaining the embedding $z_v \in \mathbb{R}^d$. Then, the decoder takes the embeddings to reconstruct some graph characteristics, such as the neighborhood $N(v)$ of node v (Figure 2.8). Popular decoders use embeddings from two nodes to predict their relationship or similarity. The similarity function S defines which type of relationship the algorithm captures, such as connection or shared local neighborhood. The optimization is performed, such as using SGD (Section 2.15) to minimize the reconstruction loss,

$$\mathcal{L} = \sum_{(u,v)\in V} l(DEC(z_u, z_v), S[u, v]), \tag{2.24}$$

so that the latent space captures the similarity between the nodes:

$$DEC(ENC(u), ENC(v)) = DEC(z_u, z_v) \approx S[u, v]. \tag{2.25}$$

Matrix factorization approaches can be used to learn a low-dimensional approximation of the similarity matrix S. For example, Laplacian eigenmaps (LE) build on the idea of spectral clustering (Section 2.4.2.3, [Belkin and Niyogi, 2003]) and use a decoder based on L2-distance between node embeddings. Intuitively, the model penalizes similar nodes in the graph whose embeddings are far apart. If the similarity is constructed satisfying the properties of the Laplacian matrix, then embeddings generated by minimizing Eq. 2.24 are identical to the spectral clustering solution. As an extension, some studies have used a dot-product in the decoder $DEC(z_u, z_v) = z_u^T z_v$, assuming this product approximates the similarity, such as the overlap of the local neighborhood [Ahmed et al., 2013, Cao et al., 2015, Ou et al., 2016]. Their loss function can be minimized using factorization algorithms, such as SVD, which is why they are called matrix-factorization approaches.

Li et al. [2021], Nelson et al. [2019], Su et al. [2020] along with Jaiswar et al. [2022] give an excellent

overview of (mostly shallow) **biological network embedding techniques and applications**. For instance, MuNK outperforms traditional approaches in terms of time and performance in the task of network alignment comparing networks across species and finding correspondence of proteins [Fan et al., 2017a]. Another example is OhmNet which learns embeddings to predict tissue-specific gene functions in a multi-layer hierarchical network, where each layer represents the interaction of genes in different tissues [Zitnik and Leskovec, 2017]. Embeddings of Mashup are obtained by applying matrix factorization method on a diffused network and used in the prediction of protein function [Cho et al., 2016].

2.4.3.1 Random walk-based shallow embeddings

While DeepWalk [Perozzi et al., 2014], as an example of a random walk-based embedding method, also uses the notion of the inner-product decoder as above, it uses a stochastic rather than a deterministic measure of node similarity (e.g., the adjacency matrix or Laplacian). Specifically, the embedding is optimized to produce similar embeddings if two nodes co-occur on random walks. The goal of the decoder is to reconstruct the probability of visiting the node v starting from node u on a T-long random walk:

$$\mathrm{Dec}(z_u, z_v) \approx Dec(z_u, z_v) \approx \textsc{Dec}(z_u, z_v) \approx P_T(v|u). \tag{2.26}$$

DeepWalk starts by generating random walks, which are a succession of random nodes across the network in predefined lengths. Then, these node sequences are input to a shallow neural network to predict a target node based on the surrounding nodes [Mikolov et al., 2013a,b]. Finally, after training, the embeddings of each node are the weights of the hidden layer of the neural network, which can be used for subsequent tasks. This network embedding approach and others, such as node2vec [Grover and Leskovec, 2016], were inspired by the natural language processing field. In word2vec, the concept is to create word embeddings that reflect the semantic relationship. Good reviews detailing the state-of-the-art in network embedding are e.g. Zhang et al. [2018] and Amara et al. [2021] conducted. In Section 2.6, we use random walks over a multi-omic network to learn node embeddings in order to characterize a node in its multi-model context of tissues, genes, genetic predisposition, and phenotypes.

2.4.3.2 Relationship of network embedding methods and matrix factorization

This separation into random walk-based and matrix factorization methods might be arbitrary since Qiu et al. [2018] showed that many network embedding frameworks are closely related to and can indeed be unified into matrix factorization frameworks with closed forms. Large-scale Information Network Embedding - LINE [Tang et al., 2015b] can be understood as a special case of DeepWalk [Perozzi et al., 2014], which performs a low-rank transformation of the normalized graph Laplacian. Furthermore, Predictive Text Embedding PTE [Tang et al., 2015a] jointly factorizes multiple graph Laplacians. The network embeddings provided by node2vec [Grover and Leskovec, 2016] are related to the stationary distribution and transition probability tensor of a second-order random walk.

2.4.4 Relational prediction in Knowledge Graphs (KGs)

Another important task is predicting links in a graph. Specifically, we will be dealing with graphs containing relations of multiple types, also known as knowledge graphs (KGs) (Figure 2.9). While ontologies and semantic databases have existed before, the term KG was coined by Google in 2012 [Singhal, 2012]. KGs have since been formally described by others, such as Ehrlinger and Wöß [2016], referring to an abstract framework to structure information to derive new knowledge by applying a reasoner. The reasoner infers missing links based on the information already present in the database, known as Knowledge Graph Completing (KGC), and relies on explicitly modeling relations between entities [Nickel et al., 2016]. Classical KGs include Freebase, Wikipedia, DBpedia, YAGO, and the Google Knowledge Graph [Färber et al., 2017, Krause et al., 2016, Mika et al., 2014, Paulheim, 2017].

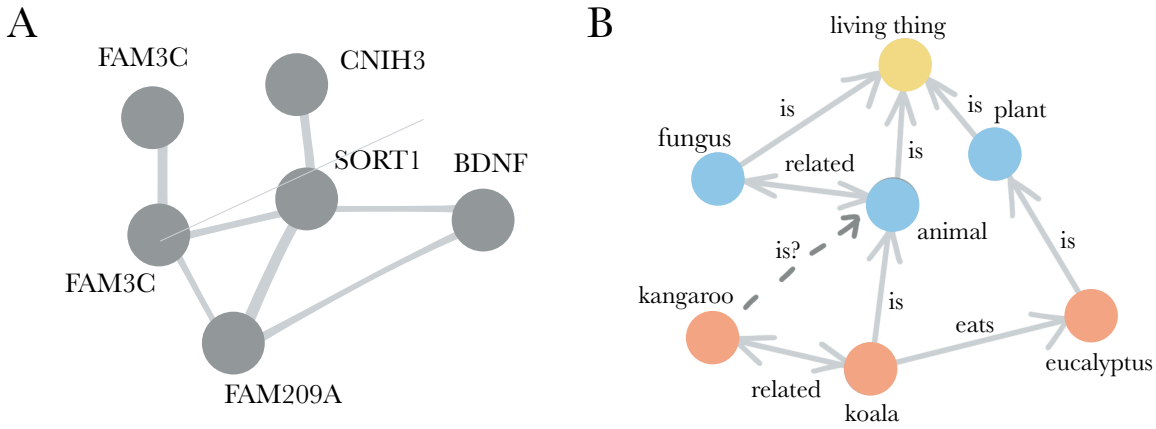


Figure 2.9: **Different types of networks:** A) Undirected homogenous network detailing PPI between proteins. B) Representation of knowledge in a directed heterogeneous network, where the Knowledge Graph (KG) is given as triplets $\langle h, r, t \rangle$, such as $\langle$koala, eats, eucalyptus$\rangle$. Different node types and relation types exist. Knowledge Graph Completion (KGC) is the prediction of missing but true links given a query $p(?|h, r)$, such as $p(?|kangaroo, is)$.

The main distinction between KGs and other graphs discussed previously is the modeling of the specific relations between entities. KGs are given by $KG = (V, R)$, with V as the set of vertices or entities and R as the type of relationship. We can then interpret that a particular fact or relation $r \in R$ holds between head node $h \in V$ and tail node $t \in V$, which is expressed as triples of the form $\langle h, r, t \rangle$. In the task of completing KG known as KGC, the meaning and context of the entities and the relationships within a graph are used to predict missing triplets or facts to draw inferences and conclusions. Questions are answered in the form of a query $p(?|h, r)$ to make informed decisions about the possibility of t, based on the existing knowledge in the graph.

KGs in **biology** include the previously introduced biological networks of BioGRID, PathwayCommons and PrimeKG (Section 2.1.5). For instance, BioGRID elucidates the indirect and direct physical interaction between proteins as PPIs. PathwayCommons further includes information on the involvement of biological entities in cellular pathways. Lastly, PrimeKG accumulates the relationship of proteins, functions, diseases, and drugs. Other such compiled resources include UniProt [Consortium, 2022], Gene Ontology [Aleksander et al., 2023, Ashburner et al., 2000], DrugBank [Law et al., 2014] and many others. Further references and reviews of databases are given in Bonner et al. [2022], Lam et al. [2023], Mohamed et al. [2021]. KGC methods have

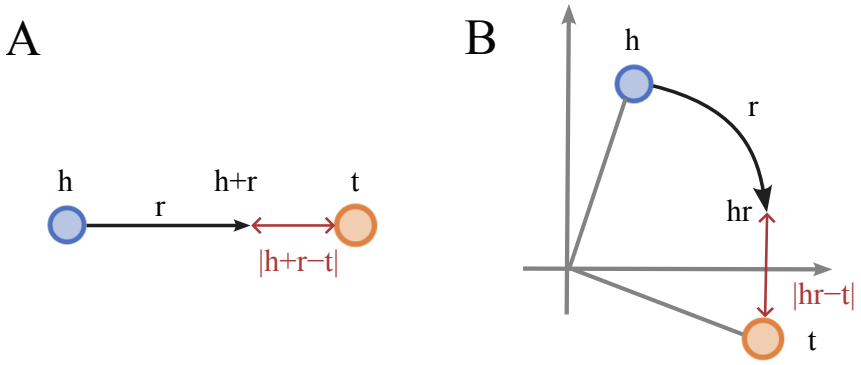


Figure 2.10: **Illustration of Knowledge Graph Embedding (KGE) methods:** A) While in the TransE model, the tail node t is modeled as a linear transformation of the head node h by the relation r, B) in RotatE, t is a rotation of h by r in complex space. The discrepancy between the prediction and the actual t is minimized during training to generate a meaningful embedding space (adapted from Sun et al. [2019]).

gained rapid development as well as improvement, and so has the incentive to use these methods with origins in computer science in the prediction of missing yet true links in the biological domain. Applications are predicting the interactions of drugs and targets [Mohamed et al., 2019], polypharmaceutical side effects [Zitnik et al., 2018] or tissue-specific protein functions [Zitnik and Leskovec, 2017].

There are several different approaches for predicting links, including two groups of methods using node representations and another using path representations. Among the embedding-based methods are Knowledge Graph Embeddings (KGE), which expand on the node embedding approaches above, and GCN-based methods. Path-based methods use message-passing neural networks to derive embeddings representing paths instead of nodes.

2.4.5 Relational prediction with KG Embedding methods

The prediction of multi-relations can also be regarded as a reconstruction task using the described encoder-decoder framework. Instead of only considering the embeddings of two nodes, h and t, the **decoder** now also considers a relation type r. For one of the simplest models, TransE [Bordes et al., 2013] (Figure 2.10A), the decoder is defined as:

$$DEC(h,r,t) = -||z_h + r_\rho - z_t||, \tag{2.27}$$

with $r \in \mathbb{R}^d$ as a learnable vector of the relation $r \in R$ and usually L1 or L2 norm. The decoder outputs the likelihood that the edge r between h and t exists and is proportional to the distance between the embeddings of t and h after *translation* according to the relation embedding of r_ρ. The embeddings with the RESCAL model are learned by minimizing the **loss function**

$$\begin{aligned}\mathcal{L} &= \sum_{h\in V}\sum_{t\in V}\sum_{r\in R} ||DEC(h,r,t) - A[h,r,t]||^2 \\ &= \sum_{h\in V}\sum_{t\in V}\sum_{r\in R} ||-||z_h + r_\rho - z_t|| - A[h,r,t]||^2,\end{aligned}$$

where the multirelational adjacency graph is given by $A \in R^{|V|\times|R|\times|V|}$, which reconstructs the local multirelational neighborhood information. Minimizing this loss function is not trivial since the nested sum is very expensive to compute. Furthermore, the MSE loss is unsuitable for reconstructing the binary adjacency matrix, which is closer to a classification case. In practice, the loss function can be the cross-entropy loss derived from logistic regression (Eq. 2.8). The log-likelihood of a correct prediction is given by $\log(\sigma(DEC(z_h, r, z_t)))$ and of an incorrect prediction is given by $\mathbb{E}_{v_n} \sim p_{t,h(V)}[\log(\sigma(-(DEC(z_h), r, z_t)))]$, with the latter approximated from a set of negative samples $\mathcal{P}_{n,h}$ sampled from $\mathcal{P}_{n,h(V)}$:

$$\mathcal{L} = \sum_{(h,r,t)\in\mathcal{E}} -\log(\sigma(DEC(z_h, r, z_t))) - \sum_{t_n\in\mathcal{P}_{n,h}} [\log(-DEC(z_u), r, z_{t_n})))]. \tag{2.28}$$

A *hinge loss* can also be applied directly using the scores from the decoder. The objective is to produce higher scores for true pairs than for false pairs. If all scores for true pairs are indeed higher by at least a margin Δ, then the loss will be 0:

$$\mathcal{L} = \sum_{(h,r,t)\in\mathcal{E}} \sum_{t_n\in\mathcal{P}_{n,h}} \max(0, -DEC(z_h, r, z_t) + DEC(z_h, r, z_{t'}) + \Delta) \tag{2.29}$$

This *contrastive learning* scheme allows the model to learn similarities between pairs from the same class and dissimilarities between pairs from different classes. In practice, negative triplets are sampled by either perturbing t or h, yielding $\langle h, r, t'\rangle$ and $\langle h', r, t\rangle$, respectively [Galárraga et al., 2013]. Sampling good negative samples is essential for all KG reasoning methods to rank missing yet true links highly.

Table 2.1: Scoring functions (i.e., decoders) used in KGE models and the logical patterns that can be captured.

Model	Scoring Function/ Decoder	Symmetry	Anti symmetry	Inversion	Composition	Reference
TransE	$-\|\|z_h + r - z_t\|\|$		✓	✓	✓	Bordes et al. [2013]
TransX	$-\|\|g_{r,1}(z_h) + r - g_{r,2}(z_t)\|\|$		✓			Wang et al. [2014] Nguyen et al. [2016a] Ji et al. [2015]
DistMult	$\langle z_h, r, z_t\rangle$	✓				Yang et al. [2015]
ComplEx	$\text{Re}(\langle z_h, r, z_t\rangle)$	✓	✓	✓		Trouillon et al. [2016]
RotatE	$-\|\|h \circ r - t\|\|$	✓	✓	✓	✓	Sun et al. [2019]

Many other KGEs exist that differ mostly in their decoders (Table 2.1). Further extensions of TransE use different translation schemes, collectively referred to as TransX [Ji et al., 2015, Nguyen et al., 2016a, Wang et al., 2014]. Here, $g_{r,i}$ depends on the relation r, which is trained to transform the entity embeddings before the translation operation. For example, TransH defines the decoder as $-\|(z_h - w_r^T z_h w_r) + r - (z_h - w_r^T z_t w_r)\|$. Furthermore, the scoring function of DistMult is the dot product between the vector embeddings of h, r, and t [Yang et al., 2015]. In contrast, RotatE uses rotation in the complex plane [Sun et al., 2019] (Figure 2.10B).

KG reasoning methods can learn a range of different logical patterns in relations, including symmetry, inversion, and compositionality. Due to the linear translation of embeddings, TransE can model hierarchical structures such as those found in tissue, disease, or Gene Ontology. While DistMult only models symmetrical relationships where the scores for triplets $\langle h, r, t\rangle$ and $\langle t, r, h\rangle$ are the same, ComplEX represents the entities as complex conjugates and is capable of modelling

asymmetrical relations. RotatE on the other hand, models a range of relations. Each model learns specific patterns due to its decoder (Table 2.1). **In biology**, Mohamed et al. [2021] have examined the capabilities of DistMult and ComplEx alongside other KGE methods in three tasks of drug target interaction, polypharmacy and tissue gene functional prediction.

2.4.6 Relational prediction with Graph Neural Networks

GCNs were introduced in Section 2.3.4.2 and can be considered convolutions on the graph structure that incorporate node features to generate node embeddings [Kipf and Welling, 2017]. However, another way to understand GNNs is using the general *message passing neural network* (MPNN) framework consisting of three main functions: MESSAGE, AGGREGATE, and UPDATE function [Gilmer et al., 2017]. We will review the basics of message passing before moving on to the specific application of link prediction with GNNs.

2.4.6.1 Fundamentals of MPNNs

In MPNNs, messages are propagated between nodes in a graph, with node embeddings primarily updated by aggregating messages from neighboring nodes and the current node. At the start, each node has an initial embedding $h^{(0)}$, which can be randomly initialized, or are the node features themselves. Then, in each message passing step (i.e., at each layer k), the initial or hidden embedding $h_u^{(k-1)}$ of each node u is updated with the aggregated messages from the node's neighborhood $N(u)$. Specifically, three components are key in the message passing framework: 1) the MESSAGE operator generates messages h_v for each node in the neighborhood $N(u)$, then 2) the AGGREGATE operator collects the embeddings of each node over the edges and combines them into a single unified signal $m_{N(u)}^{(k-1)}$ and finally 3) the UPDATE operator renews the current node $h_u^{(k)}$ using its state from the layer before $h_u^{(k-1)}$ in conjunction with the unified messages from the neighbors $m_{N(u)}^{(k-1)}$:

$$\begin{aligned} h_u^{(k)} &= \text{UPDATE}(h_u^{(k-1)}, \text{AGGREGATE}(\text{MESSAGE}(\{h_v^{(k-1)}, \forall v \in N(u)\}))) \\ &= \text{UPDATE}(h_u^{(k-1)}, \text{AGGREGATE}(m_v^{(k)}, \forall v \in N(u))) \\ &= \text{UPDATE}(h_u^{(k-1)}, m_{N(u)}^{(k-1)}) \end{aligned} \tag{2.30}$$

This process is performed iteratively over multiple layers, with each layer corresponding to a single message-passing step. The notion is that after k steps of message passing, the node embeddings capture information about all features within their k-hop neighborhood. The generalized formation of an MPNN allows the understanding of a wide range of GNNs with different MESSAGE, AGGREGATE, and UPDATE functions.

MESSAGE functions generate the messages that are to be passed through the edges to neighboring nodes. The message of one node v takes the form $m_v = \text{MESSAGE}^{(k)}(h_v^{(k-1)})$. In the most simple form, the function is an identity matrix so that the messages correspond to the node's embedding $m_v = I^k h_v^{(k-1)}$. Another example of a generated message is the multiplication of the node's embedding with a weight matrix $m_v = W^{(k)} h_v^{(k-1)}$.

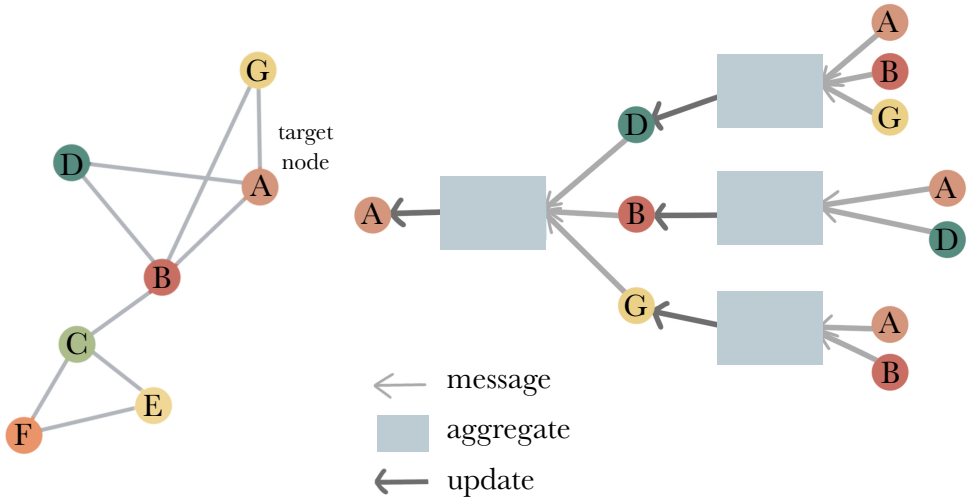


Figure 2.11: **Message passing neural network consisting of the the message, aggregation and update function:** Illustration of how the target node's A embedding is updated by the messages that are passed and aggregated from the nodes in the neighborhood N_A which in turn aggregate messages from their neighbors. This example shows a 2-layer message passing model. (Adapted from Hamilton [2020])

AGGREGATE functions summarize the information from the nodes of the graph neighborhood and has been the object of research and development. Their expressiveness differs in distinguishing nodes in different local graph structures [Xu et al., 2019] and empirical evidence found that the choice of the aggregation operator is crucial to the performance of the GNN in different tasks [Hamilton et al., 2017a, You et al., 2020]. Aggregation function include sum, max, mean, and symmetric normalization [Kipf and Welling, 2017] (Table 2.2). For instance, while the mean aggregator learns the neighborhood distribution, the max can identify the most striking node, and the sum captures information such as the node degree. However, there are scenarios where each of them performs poorly in differentiating node neighborhoods [Corso et al., 2020]. Principal Neighborhood Aggregation (PNA) uses multiple such aggregators and learns how to scale and balance the influences of each in a degree-based way for optimal learning [Corso et al., 2020]. Another important aggregation function is based on attention, weighting each neighbor's influence on the aggregation [Vaswani et al., 2017], yielding graph attention networks [Veličković et al., 2018]. The rational is that not every node is equally important and the attention mechanism allows to focus on the small yet informative part that can be learned through training. GraphSAGE introduces another complex transformation of the neighbor features through am MLP [Hamilton et al., 2017a].

UPDATE functions can be as simple as a linear combination of the neighbors and self-embedding but can also take more complicated forms (Table 2.3), especially to combat the phenomenon of over-smoothing. This happens when the number of layers is increased, but the information on local structures is lost simultaneously when the embeddings approach a near-uniform distribution. In other words, node-specific information is "washed out" with increasing iterations of message passing, when the node neighbors dominate the updated node representation. Here, skip connections can help, by concatenating the embeddings of the base update function $\text{UPDATE}_{base}(h_u^{(k-1)}, m_{N(u)})$ with the embedding of a node from the previous layer $h_u^{(k-1)}$ [Hamil-

Table 2.2: Aggregation functions of MPNN framework detailing how the messages from the neighbors are combined into $m_{N(u)}$, ranging from a simple sum, mean to more complex functions based on attention or the scaling of different aggregators.

Name	Aggregation functions	Reference
Sum	$m_{N(u)} = \sum_{v \in N(u)} h_v$	
Max	$m_{N(u)} = \max_{v \in N(u)} h_v$	
Average	$m_{N(u)} = \frac{\sum_{v \in N(u)} h_v}{\lvert N(u) \rvert}$	
Symmetric Normalization	$m_{N(u)} = \sum_{v \in N(u)} \frac{h_v}{\sqrt{\lvert N(u) \rvert \lvert N(v) \rvert}}$	Kipf and Welling [2017]
Set Aggregator	$h_u^{(k)} = MLP_\theta(\sum_{v \in \mathcal{N}(u)} MLP_\rho(h_v))$	Zaheer et al. [2017]
Neighborhood Attention	$m_{N(u)} = \sum_{v \in \mathcal{N}(u)} \alpha_{u,v} h_v$	Veličković et al. [2018] Vaswani et al. [2017] ,
Principal Neighborhood Aggregation	multiple aggregators with degree-scalers	Corso et al. [2020]

ton et al., 2017b]. Next, interpolation can be used which element-wise multiplicates the output from the base update with α_1 and the previous representation with α_2 [Pham et al., 2017]. It can be understood as an interpolation, as $\alpha_2 = 1 - \alpha_1 \in [0, 1]^d$ where the embeddings are balanced, with the exact parameters e.g. learned during training. Jumping knowledge connections concatenate not only the information from the previous layers, but for the final layer the embeddings from all previous layers are utilized [Xu et al., 2018]. Further, there are gated updates in gated GNNs [Li et al., 2017b] or edge and graph-level features that can be leveraged in Battaglia et al. [2018].

Table 2.3: Update functions of MPNN framework detailing how the aggregated messages from the neighbors of the current layer $m_{N(u)}$ and self-embedding of the last layer $h_u^{(k-1)}$ are used to update the current embedding $h_u^{(k)}$. Good update functions combat the issue of over-smoothing.

Name	Update functions	Ref
Sum	$h_u^{(k)} = h_u^{(k-1)} + m_{N(u)}$	
Skip connection	$h_u^{(k)} = \text{UPDATE}_{base}(h_u^{(k-1)}, m_{N(u)}) \oplus h_u^{(k-1)}$	Hamilton et al. [2017b]
Interpolation	$h_u^{(k)} = \alpha_1 \circ \text{UPDATE}_{base}(h_u^{(k-1)}, m_{N(u)}) + \alpha_2 \odot h_u^{(k-1)}$	Pham et al. [2017]
Gated updates	$h_u^{(k)} = GRU(h_u^{(k-1)}, m_{N(u)}^{(k)})$	Li et al. [2017b]
Residual connections	$z_u = f_{JK}(h_u^{(0)} \oplus h_u^{(1)} \oplus ... \oplus h_u^{(K)})$	Xu et al. [2018]

2.4.6.1.1 Important properties

of the MPNN are *permutation invariance* of the graph-level representation and *permutation equivariance* of node representations to node order. While the former means that the function is independent of the arbitrary ordering of the rows and columns of the adjacency matrix, the latter refers to the fact that the output is permuted consistently to the permutation of the adjacency matrix. These two properties must be satisfied for graph learning. Non-linearity, such as ReLU or Sigmoid, adds to the expressiveness of the network and can be added in the MESSAGE or AGGREGATE function.

2.4.6.2 Relational Prediction with GCNs

For the specific task of link prediction using node embeddings, we now consider more complex GNN encoders, as in Section 2.4.5. There is the challenge of accommodating different relation types. For example, the relational graph convolutional network (R-GCN) separately aggregates the information for each relation type $\rho \in R$, using a different transformation matrix W for each relation type and edge direction [Schlichtkrull et al., 2017], where the AGGREGATE function summarizing the incoming messages from neighboring nodes takes the form:

$$m_{N(u)}^{(k)} = \sum_{(\rho \in R)} \sum_{v \in N(u)} \frac{W_\rho h_v}{f_n(N(u), N(v))} \tag{2.31}$$

Another improvement made to combat the drastic increase in parameters is the sharing of parameters. The model learns an embedding for each relation and a shared tensor across all relations. Like with KGE, the decoder takes two node embeddings and produces a score for a potential edge using the DistMult factorization as a scoring function, then the loss is calculated with for each positive and synthetically generated negative. Other methods based on GCNs include Variational Graph Auto-Encoders [Kipf and Welling, 2016] and Composition-based Multi-Relational Graph Convolutional (CompGCN) Networks [Vashishth et al., 2020]. GCN approaches for link prediction exhibit improved predictive performance compared to KGEs due to the complex GNN encoder.

GNN encoders and R-GCN specifically for relational prediction has already found application in **computational biology**, such as the previously mentioned multi-drug side effect prediction [Zitnik et al., 2018]. Another recent application is the repurposing of non-pharmacological interventions (NPI) in Alzheimer's disease (AD) [Xiao et al., 2024]. The authors constructed a AD specific KG and evaluated KGE methods alongside R-GCN and CompGCN in the retrieval of interventions to prevent AD belonging to the category of complementary and integrative health as well as dietary supplements and found the best performance for R-GCN, followed closely by one of the simplest KGE method, TransE.

However, one drawback of all embedding based methods is that they are limited to predicting links within the training graph (known as a transductive setting as opppossed to the inductive setting). Since these methods rely on learning embeddings for each entity, all entities must be known at training. Further, interpretability on node embeddings could be more robust, making it difficult to understand why a given prediction was made based on two embeddings. These two disadvantages are lessened with path-based reasoning techniques.

2.4.6.3 Relational prediction with path-based reasoning techniques

Instead of using node representation for the prediction of missing links, this class of methods focuses on path representations. While some methods are based on the principals of reinforcement learning generating relational paths from the query to the answer node, such as DeepPath, MINERVA, RNNLogic and M-Walk [Das et al., 2018, Qu et al., 2021, Shen et al., 2018, Xiong et al., 2017], RED-GNN and Neural Bellman-Ford network (NBFNet) were at the forefront to learn path representations with GNNs [Zhang and Yao, 2022, Zhu et al., 2021]. These methods combine the advantages of traditional methods based on paths with the high expressiveness of

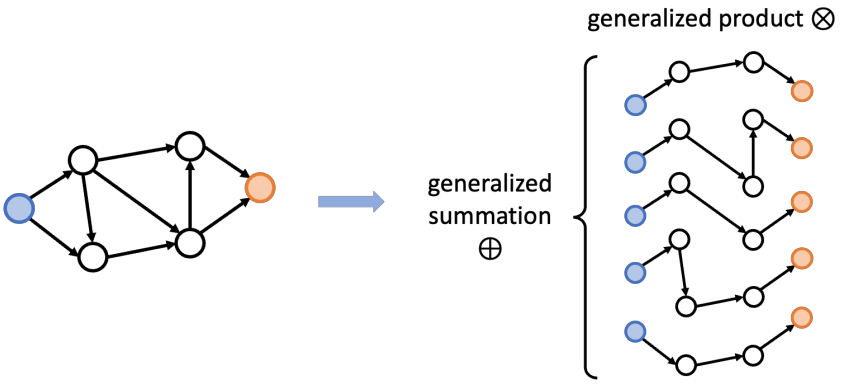


Figure 2.12: **NBFNet learns the path representation:** of the source node (blue) - target node (orange) pair as the generalized sum of all paths, where each path is the generalized product. (Adapted from and curtesy of Zhaocheng Zhu)

GNNs to automatically extract important features derived from the local neighborhood, yielding high model performance and interpretability alongside the capability of inductive reasoning. **In biology**, GNN path representation learning frameworks have been applied to predict synthetic lethality gene pairs when perturbed will lead to cell death, which is useful in the treatment of cancer by selectively killing cancer without harming normal cells [Zhang et al., 2023].

NBFNet computes a representation for pairs of nodes $h_q(h,t)$ by considering all paths between h and t for the relational prediction q. Importantly, NBFNet learns the likelihood from the entity t tail conditionally given a query consisting of the query node h and the query relation r, as $p(t|h,r) = \sigma(f(h_q(h,t)))$, where $f(.)$ is a feed-forward neural network and $\sigma(.)$ is the sigmoid function. NBFNet also augments $\langle h,r,t\rangle$ with the reverse triplet $\langle t,r^{-1},h\rangle$. Given that in a similar way, the conditional likelihood of the head entity h can be formulated by $p(h|t,r^{-1}) = \sigma(f(h_{q^{-1}}(t,h)))$.

Like before, the negative log-likelihood of the loss function is minimized during training to penalize negatively predicted relations for true positive triplets and vice versa. While positive samples are given by the triplets of the graph, negative samples must be generated according to partial completeness assumption [Galárraga et al., 2013] by corrupting either the tail or head in the positive triplet $\langle h,r,t\rangle$, yielding $\langle h',r,t\rangle$ or $\langle h,r,t'\rangle$. For simplicity, the negative sample is represented as $\langle h',r,t'\rangle$. The loss function is given as:

$$\mathcal{L} = -\log p(h,r,t) - \sum_{i=1}^{n} \frac{1}{n} \log(1 - p(h'_i,r,t'_i)), \tag{2.32}$$

where n is the number of negative samples for each positive sample, and (h'_i,r,t'_i) presents the i^{th} negative sample.

The authors of NBFNet define the pair representation h_r as the generalized sum of all paths, where each path is the generalized product of edge representations in the path (Figure 2.12). As detailed in Section 2.4.2, many successful network analysis algorithms rely on path metrics, such as the Katz index, personalized PageRank, or Bellman-Ford algorithm. The authors show that these algorithms are special cases of the pair representation with different summation and

multiplication operators. NBFNet has three components: the INDICATOR function initializes the representation at each node, the MESSAGE function learns the multiplication, and the AGGREGATE function is the summation operator.

The performance of models is evaluated by computing the rank of positive triplets $\langle h, r, t\rangle$ found in the test set against all negative triplets $\langle h, r, t'\rangle$ that do not appear in the KG, adhering to the filtered ranking protocol proposed by Bordes et al. [2013]. For the reverse relation r^{-1}, positive triplets in the form $\langle t, r, h\rangle$ are evaluated against corresponding negative triplets $\langle t, r^{-1}, h\rangle$. This method provides a rigorous evaluation, requiring the model to assign high ranks to positive samples. The evaluation metrics include mean rank (MR), mean reciprocal rank (MRR), and hits at k (Hits@k). The MR metric is determined by averaging the ranks q of positive samples among the negatives, where lower values indicate better performance, with an optimal value of 1.

$$MR = \frac{1}{|Q|}\sum_{q\in Q} q \tag{2.33}$$

MRR is computed as the average of the reciprocal ranks, $\frac{1}{q}$, which reduces the impact of outliers, such as few triplets ranked very low. MRR values fall within the range $[0, 1]$, where higher values indicate better model performance.

$$MRR = \frac{1}{|Q|}\sum_{q\in Q} \frac{1}{q} \tag{2.34}$$

The Hits@k metric measures the likelihood that the correct predictions appear within the top k ranked triplets. It can also be interpreted as the proportion of positive triplets present in the top k predictions, making it equivalent to Recall@k. Similar to MRR, its values lie within the range $[0, 1]$, with higher values indicating higher model performance.

$$H@k = \frac{\big|\{q \in Q : q < k\}\big|}{\big|Q\big|} \tag{2.35}$$

Model selection during training is usually achieved on the highest validation MRR.

2.4.6.3.1 Framework NBFNet

is constructed such that different MESSAGE (TransE, DistMult, and RotatE) and AGGREGATE (sum, mean, max, and principal neighborhood aggregation [PNA]) functions can be used [Corso et al., 2020]. Furthermore, the number of layers and the dimension of the hidden layers can be tuned. NBFNet is instantiated with six layers with 32 hidden units, ReLU as the activation function, and the feed-forward neural network $f(.)$ set to two-layer MLPs with 32 hidden units.

2.5 Statistical learning framework for multi-omics network inference using KiMONo to prioritize key regulatory factors in MDD

In this Section, I will present a framework for multi-omic integration based on network inferences approaches that utilize regularized regression. For a holistic understanding of the complex interplay of biological entities (Section 1.4), we create a condition-specific network which is derived from patient cohort data. This Section is based for the most part on my following second-author publication [Ogris et al., 2021]. In this publication, jointly with Christoph Ogris as project lead, we developed a network inference multi-omic integration method, named KiMONo. My unique contribution was the application of KiMONo on the data from a major depressive disorder (MDD) patient cohort, and the systematic comparison of the results to a state-of-the-art two-omic integration method (expression quantitative trait loci [eQTL]). The robustness analysis of the method under increasing noise and decreasing sample sizes, as well as its application to cancer profiling in The Cancer Genome Atlas [Weinstein, 2013], was performed by Christoph Ogris, and therefore is not reported in this section. Janine Knauer-Arloth, Annalisa Marsico and Nikola Müller provided supervision and scientific discussion.

- Ogris C., **Hu, Y.**, Arloth, J., & Müller N. **Versatile knowledge guided network inference method for prioritizing key regulatory factors in multi-omics data**. *Scientific Reports, 2021*

2.5.1 Network inference

Our contribution towards integrating multi-omic data lies in the construction of a framework, named KiMONo (knowledge-guided multi-omics network inference) to uncover statistical dependencies across different omic data types (Section 1.4) and assemble their interplay into a network (Fig 2.13). Here, the preselection of features before inference allowed us to reduce the inference complexity, for which we leveraged prior knowledge of biological mechanisms. This prior knowledge was based on existing experimentally verified interactions such as PPIs (Section 2.1.5). We used the biological interactions as undirected edges. Next, linear multivariate regularized regression models (Section 2.2.3) were used to retain statistical dependencies with sparse-group LASSO penalty [Simon et al., 2013] based on the principal of FS (Section 1.4) to minimize the following error term:

$$\min_{\beta} \frac{1}{2}||y - \sum_{l=1}^{m} X^{(l)}\beta^{(l)}||_2^2 + (1-\alpha)\lambda \sum_{l=1}^{m} X^{(l)}\sqrt{p_{(l)}}||\beta^{(l)}||_2 + \alpha\lambda||\beta||_1,$$

where α is the strength of regularization $[0, 1]$, and $\beta^{(l)}$, $p^{(l)}$, and $X(l)$ are the coefficients, length of the coefficients, and matrix of the lth group of omics. Using the sparse-group LASSO, two types of penalties were used to induce two types of sparsity. The first acted on all groups (i.e., omic layers) and is called the "group-wise" sparsity. The second acted on each feature within a group (i.e., omic layer). This approach allowed for a flexible framework to determine

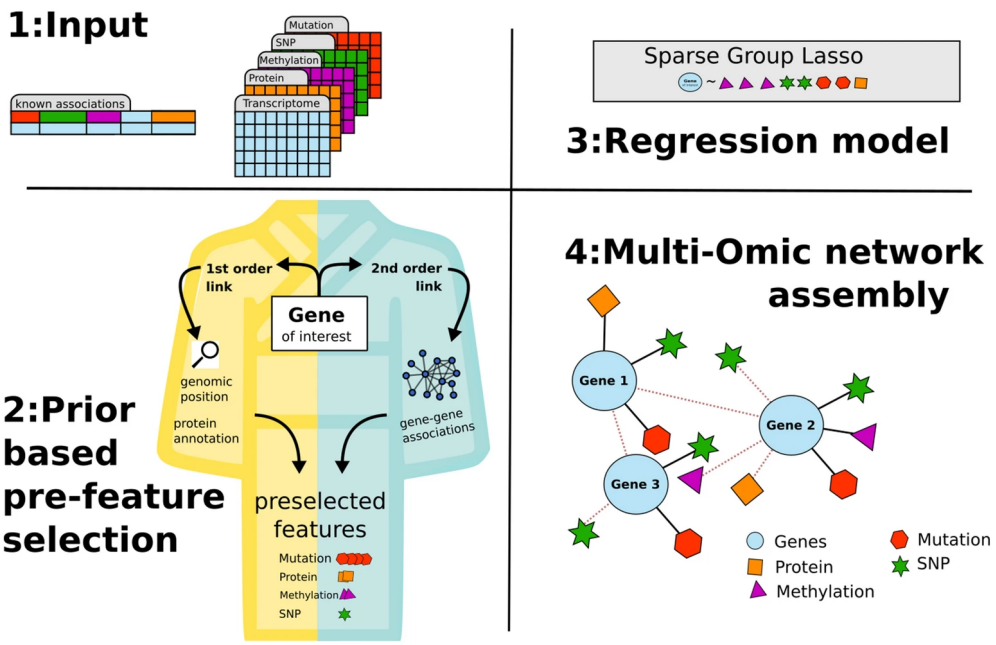


Figure 2.13: **Steps in the workflow of KiMONo:** our method integrates prior knowledge in the form of known biological interactions and multiple omic data sources to construct a comprehensive multi-omic network of interconnected features. 1: KiMONo accepts input data that can comprise various types of omic data along with prior biological knowledge. This prior knowledge is provided as a list of known associations between the input features. 2: Using prior knowledge, KiMONo performs a prior-based preliminary selection of omic features and creates an input matrix X for each gene. 3: KiMONo builds a regression model for each gene using a sparse-group LASSO regression approach. In this model, gene expression is the dependent variable (y), and the previously selected matrix X is the input. 4: All gene models are combined to generate a multi-level omic network. This network encompasses features from all input sources, representing them as nodes, and establishes links between them based on the non-negative regression coefficients derived from the models. (From Ogris et al. [2021]

the importance of each omic group and its component features. In the final step, all retained features with a coefficient $|\beta| > 0.01$ were connected to their dependent variable y to form a multi-omic network.

The performance of each gene model was assessed using the R^2 measurement, which represents the percentage of variance in the regressed variable (i.e., gene expression) that can be explained by the regressors (i.e., multi-omic features). It was calculated as $R^2 = ESS/TSS$, where the explained sum of squares is given by $ESS = \sum(\tilde{y} - \overline{y})^2$, and the total sum of squares by $TSS = \sum(\tilde{y} - y)^2$. Here, $\tilde{y}$, $\overline{y}$, and y are the predicted, mean, and measured (true) gene expression. Thus, $R^2 = [0, 1]$ measures the goodness of fit for every gene model, with values closer to 1 indicating a better fit (Section 2.2.2). The tool can be found at `https://github.com/cellmapslab/kimono`).

2.5.2 Network analysis

In our analysis, we treated all retained features from multiple levels as undirected edges. This transformation allowed us to generalize the complex directed network into a more simplified

representation resembling a single-layer association network. To assess whether the generalized network structure conformed to the typical characteristics of biological networks, such as being scale-free, we tested the goodness of fit of the distribution of edges to the power law [Simon et al., 2013]. We also used betweenness centrality as a metric to estimate the significance of nodes within this network. Centrality was calculated as $(v) = \sum \frac{\sigma_{uw}(v)}{\sigma_{uw}}$, where $\sigma_{uw}(v)$ was the shortest path between nodes u and w, passing through node v, reflecting its importance in the overall network (Section 2.4.2.1).

2.5.3 Case study: Integrating multi-omic data for a MDD patient cohort

We showcase our network integration method, KiMONo, using a cohort of patients with MDD. While efforts have been made to understand the pathomechanism of the psychiatry disease MDD, translating the findings into clinical practice has been limited [Kapur et al., 2012]. While most prior studies have focused on analyzing a single omic level, such as genome-wide association studies (GWAS) [Anderson et al., 2020, Arloth et al., 2015b], studying multiple omics is relatively new. However, comprehending multi-omic cross-talk would advance understanding of the disease phenotype. Therefore, our study focused on integrating different data modalities into a regulatory network to examine their interplay and identify important markers.

2.5.3.1 Data availability

The study cohort comprised 289 Caucasian individuals: 129 patients with MDD and 160 healthy controls. The recruitment methods and additional characterization of the MDD group have been previously described by Arloth et al. [2015b], Zannas et al. [2015]. Data for three omic layers (transcriptome, methylome, and genotype) and clinical data were collected from a subset of 107 of the 289 individuals, including 33 females and 74 males, distributed among 64 controls and 43 patients. The omic data preprocessing has been comprehensively described by Arloth et al. [2015b], Zannas et al. [2015].

2.5.3.2 Leveraging prior information

Prior information was used to perform the preselection of features and yielded the input for the regularized regression models. Here, we made a distinction between first- and second-order connections. To establish the prior knowledge of first-order connections, we used the Re-Annotator pipeline [Arloth et al., 2015a] with RefSeq build GRCh37 (hg19) of the human reference genome to annotate gene expression probes and gene symbols. In addition, we used Bismark [Felix Krueger, 2011] to realign the methylome CpG site probes and transcriptome gene symbols to specific sequence positions. Moreover, we established associations between genes and SNPs within a 10 kbp distance and methylation sites within a 500 kbp distance. Second-order connections were formed between genes using a "guilt-by-association" approach, leveraging data from the BioGRID database [Oughtred et al., 2021, Stark et al., 2006]. Furthermore, we connected genes to their associated methylation sites, forming second-order linked methylation sites. These associated genes, methylation sites and SNPs were then used as input for sparse group lasso feature selection in the regularized linear regression models.

2.5.3.3 Generation of a heterogeneous MDD network

For each of the genes, we first performed the feature pre-selection to determine all associated genes, methylation sites, SNPs of first- and second-order associations. Then, alongside with all clinical/phenotypic information of sex, age and BMI, regularized models were calculated, with the gene as response variable y and the preselected features as X. After sparse group lasso model fitting, quality cutoff were applied to models with $R^2 > 0.1$ and to features with a higher absolute coefficient $\beta > |0.02|$. To generate the multi-omic network the retained multi-omic features were connected to the response gene variable. Iteratively, this process was applied to all genes. Then, upon obtaining the multi-omic network, the centrality measured as betweenness of each node was calculated.

2.5.3.4 Comparison with a pairwise omic integration tool

The eQTLs (Section 1.4.3) are identified by testing the correlation between the expression of a gene and proximal (cis-regulation) or distal (genome-wide; trans-regulation) genetic variants. In contrast, our method implicitly calculates both pairwise and multivariate eQTLs while imposing a genomic distance cutoff for linking variants and gene expression. For the MDD dataset, we used both methods to detect eQTL and eQTM genes. We used the MatrixEQTL [Shabalin, 2012] (version 2.3) pairwise analysis tool to retrieve these genes, with 10 or 500 kbp window around the gene of interest, respectively. Moreover, we corrected for covariates on the expressed genes, considering factors such as BMI, age, sex, and diagnostic status, with a significance threshold set at FDR < 0.05. In the context of KiMONo, the identification of eQTL and eQTM genes relied on the inferred cross-layer interactions between genes, methylation sites, and SNPs. We considered results robustly inferred when the models exhibited R^2 values of ≥ 0.1 and corresponding cross-layer associations of ≥ 0.2.

2.6 ML framework for multi-modal network inference and embedding to contextualize COVID-19 genes

While previous method solves the integration of multi-omic data, this section describes a framework I developed to first integrate multi-modal data into a network, then embed the nodes into latent space for efficient analysis (Section 1.5). We applied it to contextualize COVID-19 genes in its multi-modal space using a pre-pandemic population cohort. It is based on and partly identical to the following first-author publication [Hu et al., 2022]. In this publication, I conceptualized and designed the study and the framework, supervised by Annalisa Marsico, Nikola Müller, and Janine Knauer-Arloth. Further, I implemented the method and analyzed the results with Ghalia Rehawi, supported by Nathalie Gerstner and Florian Bittner.

- **Hu, Y.**, Rehawi, G., Moyon, L., Gerstner, N., Ogris, C., Bittner, F., Marsico, A., and Mueller, N.S.. **Network embedding across multiple tissues and data modalities elucidates the context of host factors important for COVID-19 infection**. *Frontiers in Genetics, 2022*

2.6.1 Network inference and embedding

Our two-step approach for investigating the similarities between multi-modal data initially constructs a multi-modal network and then projects the nodes into a low-dimensional continuous embedding space that captures graph topology and relationships between nodes. These vector representation of nodes enable similarity computation for efficient analysis of a complex heterogeneous network (Section 1.5.3 and background Section 2.4.3). We used KiMONo to generate the multi-omic network by statistically selecting features that contribute to predicting each gene's expression pattern (detailed in Section 2.5 above). KiMONo's FS process operates on both modality groups (e.g., genes and phenotype) and their component features. The features selected by the sparse-group LASSO model were incorporated into the network as nodes linked to the modeled gene node. We constructed a multi-modal network from all the KiMONo statistical models by connecting all the modeled features with their corresponding explanatory variables (Figure 2.14B). Stability selection was conducted over 30 runs to ensure the reliability of the selected features. Features were retained if selected in more than 70% of the runs, thus considering only robustly selected features. Default filtering steps were also applied to the inferred gene models. Connections with an R^2 value greater than 0.01 and an absolute mean β coefficient larger than 0.01 after filtering were retained to reduce noisy connections and ensure the inclusion of high-quality models.

In the second step of our ML framework, we learn the low-dimensional node embeddings of the multi-modal network using the embedding method following GeneWalk [Ietswaart et al., 2021] (Section 1.5.3 and method background 2.4.3.1). This method is based on DeepWalk [Perozzi et al., 2014], which starts by generating sequences of nodes through unbiased random walks across the network (Figure 2.14C). A random walk is a stochastic process with random variable starting at the target node $x_1 = v$ and $x_2 = u$ uniformly chosen from the neighbors of v. This process of selecting neighbors is repeated until a predefined length for a random walk is reached.

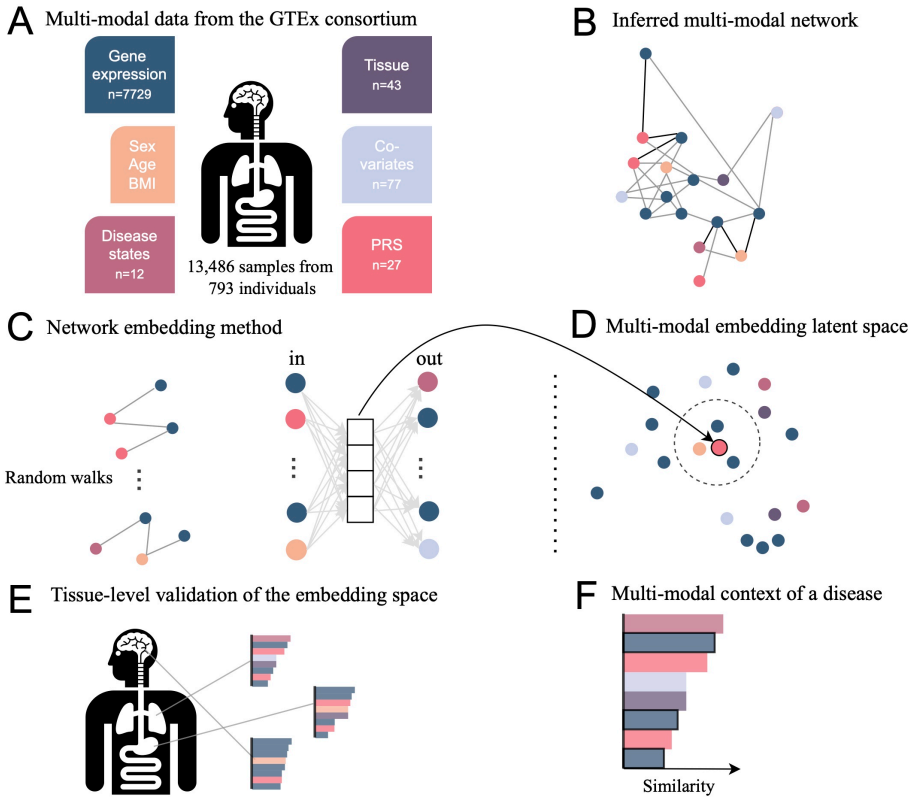


Figure 2.14: **Network inference and embedding:** A two-step ML framework comprising multi-omic network inference followed by embedding for efficient investigation in continuous low-dimensional space. A) Data was obtained from the GTEx consortium, consisting of gene expression in many tissues, phenotypic attributes (sex, age, and BMI), diagnostic information, technical and biological covariates, and polygenic risk scores (PRS) indicating a genetic predisposition to specific diseases. B) After multi-modal data integration, KiMONo inferred a comprehensive multi-tissue network. C) Then, embeddings of the obtained network are learned using an adapted GeneWalk method which is based on the results from random walks across the network. D) The embedding of a node is its weight in the hidden layer and can be explored using nearest neighbor methods, such as cosine similarity scores. The embedding space is validated using (E) tissue-specific expression patterns and (F) delineating the multi-modal context of previously identified genes essential to COVID-19. (From Hu et al. [2022])

The sequences of nodes in a random walk can be thought as analogous to the words of a sentences. Next, a neural network consisting of one layer is trained to predict the target node v given the neighboring nodes u_j in the random walk sequence. This scheme follows the SkipGram model [Mikolov et al., 2013a,b] that maximizes the co-occurrence probability on a random walk:

$$Pr(u_{i-w}, u_{i-1}, u_{i+1}, ...u_{i+w}|ENC(v_i)) = \prod_{\substack{j=i+w \\ j=i-w \\ j \neq i}} Pr(u_j|ENC(v_i)) \tag{2.36}$$

with window size w, determining the truncated length of the random walk can be adjusted to include larger or smaller neighborhoods and capture different hops of neighboring node pairs. The probability distribution of nodes is approximated using hierarchical softmax or a negative sampling scheme [Mikolov et al., 2013a,b, Perozzi et al., 2014]. After completed training, the weights of the hidden layer of the neural network represent the embeddings of nodes $ENC(v) = z_v$, corresponding to a lookup table. These embeddings represent the node's position in the low-dimensional space (Figure 2.14D and method background Section 2.4.3). We determine the proximity between two nodes in the embedding space using similarity scores, i.e. computing the cosine similarity between their embedding vectors.

The ML framework was implemented in R and Python and is freely available at `https://github.com/cellmapslab/embed_multimodalNet`.

2.6.2 Case study: Efficient contexualization COVID-19 related genes

Given the complex nature of COVID-19 and the diverse factors such as genetics, general risk, and comorbidities contributing to its different manifestations (Section 1.5.1), we aimed to develop a comprehensive multi-modal perspective, focusing mainly on genetics and comorbidities, across the whole body. We projected the multi-omic data into a joint embedding space to efficiently explore the relationships between different modalities. However, there was limited large-scale COVID-19 data encompassing clinical phenotypes, genomics, and transcriptomics across various tissues, restricting the full exploitation of multi-omics data integration methods to establish a global multi-tissue and cross-individual perspective of the disease. Therefore, we leveraged the Genotype-Tissue Expression (GTEx) consortium data spanning a population cohort that included comprehensive multi-omics across tissues and phenotyping data predating the COVID-19 outbreak [Carithers et al., 2015]. Here, we adopted a different perspective to comprehend the complexity in symptoms, affected tissues, and genetic variations on the individual level in the molecular response to COVID-19. To achieve this, we developed a novel strategy that combines network inference and embedding to inspect this pre-pandemic population data to uncover patterns. We then used this knowledge to gain a broader understanding of the significance of host factors in COVID-19, considering their known associations with other existing diseases, phenotypes, genetic variations, and gene expression across diverse tissues from GTEx.

In the first step, we computed polygenic risk scores (PRSs) given SNPs that reflected the genetic susceptibility to develop a specific disease. To achieve this, we made use of GWAS summary statistics for various diseases linked to COVID-19, including pneumological, cardiovascular, and metabolic diseases. Next, we used KiMONo to integrate gene expression across tisses, the previ-

ously calculated PRSs with phenotype and disease states (considered comorbidities for COVID-19) to construct a multi-modal network. Next, we employed a graph embedding method inspired by the DeepWalk algorithm [Perozzi et al., 2014], which utilizes shallow neural networks to generate embeddings for individual nodes. The embeddings summarize the associations between nodes in the multi-modal network as a single similarity value for each node pair, enabling efficient exploration and interpretation of the complex network. Furthermore, we contextualized genes in their multi-modal space that were identified in diverse experimental studies related to COVDI-19. The described genes originated from CRISPR Schneider et al. [2021], Wei et al. [2021], GWAS [The COVID-19 Host Genetics Initiative, 2020], patient omics data derived from blood serum and plasma [Demichev, 2021, Di et al., 2020, D'Alessandro et al., 2020, Geyer et al., 2021, Messner et al., 2020, Overmyer et al., 2021, Shen et al., 2020, Wu et al., 2021] and physical binding experiments [Gordon et al., 2020, Lee et al., 2021]. This approach enabled us to elucidate the connections between known COVID-19 disease states, genes, tissues and genetic risks, collectively called the multi-modal context. Using our statistical framework combining inferring and embedding multi-omic networks, we obtained insight surpassing traditional network statistics, providing a comprehensive understanding of known COVID-19 genes in the context of multiple modalities.

2.6.2.1 PRS computation

We used data on 984 individuals from the GTEx consortium, which included phenotypic information, gene expression, and genomic variations (SNPs) (Figure 2.14A). PRSs were used to quantify the genetic predisposition for developing specific diseases. We computed these scores using GWAS summary statistics for a range of diseases, including MDD and type II diabetes (T2D), and three studies on COVID-19 susceptibility, hospitalization and severity [The COVID-19 Host Genetics Initiative, 2020]. Supplementary Table S1 lists the GWAS source. Then, we lifted the individual-level genotype data from build GRCh38 to GRCh37/hg19 of the reference genome using the LiftOverPlink tool [Shaun Purcell, 2019] to align with the GWAS summary statistics. This process resulted in 1,119,899 SNPs that were successfully mapped and used to calculate 27 PRSs for 866 individuals. The PRS-CS tool [Ge et al., 2019] was used for PRS prediction, which implements a Bayesian regression approach with continuous shrinkage of SNP effect sizes. The method uses an external LD reference panel to take correlation among nearby SNPs into account, which was the European LD reference panel based on data from the 1000 Genomes Project [Ge et al., 2019] in our case. The determination of the global shrinkage parameter, ϕ, which is essential for adjusting effect sizes was set based on the degree of polygenicity of each disease and sample size, as it depends on the sparseness of the genetic architecture of a trait [Ge et al., 2019]: (1) Traits with large GWAS sample sizes ($\geq 250,000$), the ϕ parameter was chosen to be the value estimated from the data using a fully Bayesian approach. (2) Polygenic traits with sample sizes $<$ 250,000 and not more than 100 significant SNPs ($p \leq 5e-8$) were considered to have low polygenicity. The ϕ parameter was set to $1e-4$ for these traits. (3) Traits with sample sizes less than 250,000 and more than 100 significant SNPs ($p \leq 5e-8$) were considered highly polygenic. The ϕ parameter was set to $1e-2$ for these traits. Finally, we used PLINK2 (PLINK v2; Chang et al. [2015], Shaun Purcell [2019]) to calculate the overall risk for various diseases and traits for each individual in the GTEx dataset.

2.6.2.2 GTEx data processing

The GTEx consortium obtained gene expression measurements across various tissues and sub-tissues. To reduce sex-specific effects, we filtered the genes to exclude those on the X and Y chromosomes, in accordance to previous studies [Melé et al., 2015, Saha et al., 2017]. Mitochondrial genes, which are under different transcriptional control, were also excluded from the analysis. Further, we filtered out genes with low expression, retaining only those with a minimum of 0.1 transcripts per million (TPM) in at least 80% of samples. Furthermore, genes were considered if present in the BioGrid PPI database [Oughtred et al., 2019]. Regarding sample selection, we excluded samples originating from reproductive system tissues, such as the uterus, ovary, testis and prostate, to minimize biases between sexes and tissues with insufficient sample sizes ($n < 100$). Ultimately, out of the initial 56,200 genes 7,251 passed the filtering process, as did 44 sub-tissues from 30 different tissues (Supplementary Figure 4.1).

Technical covariates at the tissue level included sequencing platform, polymerase chain reaction-based sequencing mode, PC genotyping components, and probabilistic estimation of expression residual (PEER) factors [Stegle et al., 2010] that address technical sequencing conditions as confounding factors. For the calculation of the PEER factors, PCA is performed to decompose data variations attributed to factors such as batch effects and genotyping components while considering the phylogenetic relationship among individuals. Tissues were one-hot encoded for each gene expression sample for network inference. Cultured fibroblast samples were used as the reference in the sparse-group LASSO model due to their distinctiveness from other tissue groups. In addition, phenotypic information, including BMI, sex, age, and disease states, such as ischemic heart disease, renal failure, MDD and liver disease were binary encoded. These features constituted the input for the network inference using KiMONo. In total, the study comprised 13,486 samples from 793 individuals with existing diseases ($n = 12$), phenotypes ($n = 3$), gene expressions ($n = 7251$), tissues ($n = 44$), calculated PRSs ($n = 27$) and covariates ($n = 78$) available.

2.6.2.3 Multi-modal COVID-19 population cohort network

For each gene i, we incorporated the gene expression of its direct interaction partners as additional predictors, a fundamental aspect of the KiMONo approach. We used the BioGrid PPI database containing experimentally validated interactions to preselect gene-gene (G-G) interactions, as well as a 'all-against-all' prior for phenotypes, disease states, phenotypic information and tissues as input for each gene (no explicit prior filtering). Reverse models were computed by modeling the values of the non-gene features based on all the previously selected genes to prevent statistical bias toward edges between network nodes without prior information. Only the most important genes, ranked by their absolute β values (top 30%), were kept to ensure consistent edge magnitudes between G-G and gene-non-gene interactions.

To determine the robust similarity between a query node and other nodes, we considered a node robustly similar if it appeared in the top 1,000 most similar nodes in at least 80 of the 100 embedding runs. We then ranked the associated robust nodes of each query node based on their maximum similarity score to provide a multi-modal contextualization for the genes, allowing us to explore the relationships between COVID-19-associated genes and other nodes, such as tissues

or diseases.

We conducted a grid search to determine the optimal parameters using a smaller network, limited explicitly to brain samples, which will be named the "brain network." We tuned window size with options $w = [2, 3]$ to defining positive examples, the dimension of the embedding with options $d = [4, 8, 16, 32]$ during training. We chose the set of parameters with the highest variance in the similarity distribution among 10,000 randomly sampled nodes. The highest variance indicates that the network's nodes contain the most informative content while not overfitting the data. The determined embedding parameters were a window size of 2 and an embedding dimension of 16.

Then, to capture the variability introduced by the stochastic walk samples, we embedded the entire network 100 times, resulting in 100 distinct sets of vector embeddings.

In all 100 of these embedding spaces, we examined the relationships between nodes by identifying those with the highest cosine similarity scores to a specific query node of interest, such as a disease or comorbidity node. For each query node, we specifically extracted the top 1,000 most similar nodes based on the cosine similarity score across all 100 embedding runs.

2.6.2.4 Tissue enrichment analysis

When examining tissue enrichment, genes with a tissue-specific preference or expression would be expected to be closer to the corresponding tissue node than to the nodes representing other tissues (Figure 2.14E and F). To validate our approach, we compared the mean similarity scores of the topmost similar and least similar genes to tissue nodes across 100 embedding runs, considering different numbers of genes (50, 100, 200, 300, and 500). We validated the tissue specificity using genes with tissue-enhanced expression data in the Human Protein Atlas [Uhlen et al., 2015]. For instance, to validate the relationship with the brain, we searched the web server (proteinatlas.org) with the query "tissue_category_rna: brain; tissue enhanced AND sort_by: tissue specific score."

To validate our approach, we specifically focused on the liver and brain tissues since they had the most samples, as well as, tissue-enhanced genes in the Human Protein Atlas. To assess the enrichment of tissue-enhanced genes, we calculated the odds ratio by comparing the presence of tissue enhancement within the set of genes that were most similar to the tissue node versus the set of least similar genes. We compared the expression levels of genes within the GTEx dataset in the most similar tissue with the levels observed in other tissues to validate their tissue specificity.

2.6.2.5 The multi-modal context of COVID-19–related host factors

Datasets were collected and compiled from fourd different experimental techniques that were related to COVID-19, in order to study the multi-modal context of disease associated genes to shed light on the disease pathology. We investigated the context of genes from the following four types of lists. Our focus was on disease states, tissues, and PRS in the proximity in the embedding space of these SARS-CoV-2/COVID19–associated proteins, genes, and genetic variants. In this case, instead of selecting the top 1,000 nodes as previous, we set a threshold for the similarity

score (>0.65) to allow for a greater neighborhood in the latent space and include more non-gene nodes. This adjustment was necessary because nodes close to COVID-19 genes were primarily other gene nodes. Nodes that exceeded this threshold were represented in similarity-based graphs, visualizing the similarity of disease states, tissues and PRS to the genes identified in the literature. These graphs specifically included genes close to nodes associated with COVID-19 susceptibility, severity, and hospitalization. This approach allowed us to focus on genetic relevance for further analysis.

2.6.2.5.1 Genetical factors

The full summary statistics of COVID-19 GWAS (excluding the 23andMe data) were obtained for build GRCh38 of the human reference genome (released on June 6, 2021). Significant SNPs ($p \leq 1e-3$) were identified based on comparisons of very severe cases versus the general population (A1), hospitalization versus non-hospitalization (B1), and hospitalization versus the general population (B2) [The COVID-19 Host Genetics Initiative, 2020]. The significant SNPs were mapped to genes using release 101 of the ENSEMBL gene annotations using the knowing01 Explore software. This analysis identified 515, 663, and 475 genes for the A1, B1, and B2 comparisons, respectively.

2.6.2.5.2 Host-viral direct interactions

Physical interaction studies were conducted using ribonucleoprotein capture and immunoprecipitation methods [Gordon et al., 2020, Lee et al., 2021] to explore viral-host interactions. The ribonucleoprotein capture experiment focused on the "SARS-CoV-2 RNA interactome," which included 109 proteins [Lee et al., 2021]. The immunoprecipitation experiment used high-confidence scoring criteria, including a mass spectrometry interaction statistic (MiST) score ≤ 0.7, a significance analysis of interactome express (SAINTexpress) Bayesian false-discovery rate (BFDR) ≤ 0.05, and an average spectral count ≤ 2 [Gordon et al., 2020].

2.6.2.5.3 CRISPR screens

A set of genes was also derived from CRISPR studies that identified host factors essential for SARS-CoV-2 infection [Schneider et al., 2021, Stephenson et al., 2021]. The top 20 anti-viral and pro-viral genes were selected and ranked based on the mean z-score in the Cas0-v2 conditions [Wei et al., 2021]. Also, hits that were significant from Huh-7.5 37°C SARS-CoV-2 experiments were included [Schneider et al., 2021].

2.6.2.5.4 Previously identified multi-omic factors

Eight studies were screened for multi-omics data from patients. Regulated proteins from proteomics studies were identified using a lax significance cutoff (adjusted $p < 0.1$) unless otherwise specified due to a low hits overall. Geyer et al. [2021] provided pairwise comparisons of three

time points (first day of sampling, day of highest signal, and negative test for SARS-CoV-2) using sera from 262 in-patients. [Shen et al., 2020] provided comparisons of three groups (healthy, non-severe, and severe COVID-19) using sera from 19 individuals. [D'Alessandro et al., 2020] provided comparisons between controls and patients with varying COVID-19 severities using sera from 38 individuals, using a less strict filtering criterion of $p < 0.05$ due to the absence of information regarding multiple testing correction. [Messner et al., 2020] used sera from 104 patients with different severities in COVID-19 identifying biomarkers without applying additional filtering.

Using a discovery cohort of 33 individuals, 90 proteins differentially regulated between control and COVID-19 patient sera were extracted without applying any additional cutoff [Di et al., 2020].

In a study by [Demichev, 2021], proteomics blood plasma data from 139 inpatients were correlated with diagnostic parameters ($n = 86$) and associated with disease severity using a lax-adjusted cutoff ($p < 0.1$). Two additional multi-omic studies were included and applied rigorous cutoff criteria on the transcriptome data.The first study focused on COVID-19 patients ($n = 231$) without comorbidities and performed pairwise comparisons among asymptomatic, mild, and severe cases using serum proteomics, applying a relaxed adjusted cutoff ($p < 0.1$), and whole-blood next-generation RNA sequencing (RNA-seq), applying a stringent gene cutoff (adjusted $p < 1e - 10$) [Wu et al., 2021]. The second study used data from 128 individuals to assess the association between disease state and ICU care Overmyer et al. [2021]. They analyzed leukocyte transcriptomics (adjusted $p < 1e - 10$) data, plasma proteomics (adjusted $p < 0.1$) and the ICU $\times$ COVID-19 interaction for both omics (adjusted $p < 0.1$).

2.7 DL path-representation learning for link prediction using BioKGC for functional annotation and drug repurposing

The previous methods dealt with the integration of multi-omic data and the embedding thereof. The method of this section takes specific relationships between biological entities into account and aims at the prediction of missing links. In this Section, I will present a path-based reasoning tool, Biomedical KG Completion (BioKGC), which I developed together with the authors of the general path-based reasoning tool NBFNet. This section is based for the most part on the following first-author manuscript [Hu et al., 2024]. In the manuscripts, I conceptualized and designed the study in collaboration with Sophie Xhonneux and Zhaocheng Zhu from the Mila-Quebec AI Institute under the supervision of Annalisa Marsico and Jian Tang. I implemented the biomedical requirements, adapting NBFNet, with Sophie Xhonneux and Zhaocheng Zhu. I analyzed the results with Sophie Xhonneux, Samuele Firmani, and svitlana Oleshko with support from Maria Ulmer. Not detailed here are the further applications involving the prediction of synthetic lethality gene pairs and target prediction of lncRNA regulation performed by Svitlana Oleshko and Hui Cheng.

- **Hu, Y.**, Oleshko, S., Firmani, S., Zhu, Z., Cheng, H., Ulmer, M., Arnold, M., Colomé-Tatché, M., Tang, J., Xhonneux, S., and Marsico, A.. **Path-based reasoning for biomedical knowledge graphs with BioKGC.**, *bioarxiv, 2024*

2.7.1 BioKGC and adaptations to biomedical KGs

In order to determine the relationship between a query node (blue) and an answer (orange) node in KGs which is important for biomedical applications (Section 1.6 and method background 2.4.4), links can generally be predicted in two ways: KG node embedding and path embedding methods (Figure 2.15A). In the former, the embedding space is constrained to one-hop relations during training (by minimizing the distance between the transformation of h by r, while inference is typically performed over multi-hops (Section 2.4.5). The latter considers all multi-hop paths between two nodes during training and inference of a given length (e.g. $k = 6$) (Section 2.4.6.3). We introduce BioKGC, a path-based reasoning method adapted from NBFNet [Zhu et al., 2021] to predict links in biomedical KGs. NBFNet evaluates all potential tail entities by ranking them based on their probability of forming a valid triplet with a specified head entity and relation (the query). To consider all paths efficiently, NBFNet leverages the generalized Bellman-Ford shortest path algorithm [Baras and Theodorakopoulos, 2010] with efficient computation due to dynamic programming. It is a single source node-dependent framework that computes pairwise relationships from the source node to all target nodes using message passing (Figure 2.15B).

Specifically, the representation of a path is given by:

$$h_q^{(0)}(h,t) \leftarrow \mathbb{1}_q(h=t) \tag{2.37}$$

$$h_q^{(k)}(h,t) \leftarrow \left(\bigoplus_{(x,r,t)\in V} h_q^{(k-1)}(h,t) \otimes w_q(x,r,t) \right) \oplus h_q^{(0)}(h,t) \tag{2.38}$$

where $w_q(x, r, t)$ represents the edge $e = (x, r, t)$ with relation type r, and the boundary condition outputs 1 if $h = t$ and 0 otherwise. Thus, NBFNet computes the pair representation $h_q^{(t)}(h, t)$ for the query, consisting of the node h and the relation r, to all nodes $t \in V$ in parallel.

Fitting into the MPNN framework (Section 2.4.6.1), NBFNet can be defined as:

$$h_k^{(0)} \leftarrow \text{INDICATOR}(h, t, q) \tag{2.39}$$

$$h_k^{(t)} \leftarrow \text{AGGREGATE}\left(\{\text{MESSAGE}(h_x^{(k-1)}, w_q(x, r, t) \mid (x, r, t) \in V\} \cup \{h_t^{(0)}\}\right) \tag{2.40}$$

where MESSAGE is $\otimes$, and AGGREGATE is $\oplus$. These path representations are fed to a MLP to learn positive and negative relationships between nodes in a supervised manner, minimizing the negative log-likelihood of the loss function. More background and details can be found in Section 2.4.6.3.

Biological KGs (Section 2.1.5) differ from classical KGs in two important ways: 1) they specifically aim to predict links between a small subset of nodes, such as gene-pathway or drug-disease relations and 2) they contain more noise, with potentially false relations due to experimental errors, or irrelevant relations for the specific prediction task. Many KGC methods, such as KG embedding (Section 2.4.5), are poorly suited for this scenario due to their sensitivity to noise and uniform treatment of all triplets. Our key adaptions for BioKGC are 1) considering node-types during negative sampling and 2) utilizing a background regulatory graph (BRG) only for message passing while supervising on another set of edges. Both aspects are consistent with the specific requirements and characteristics of biomedical KGs.

Negative sampling is a key aspect of training in KGC methods. To enable good predictive performance, we must simultaneously assign a high score to true triplets (h, r, t) and a low score to false triplets, sampled by perturbing either h or t. Since some triplets are more easily distinguished as false than others, the negative samples must be sufficiently difficult for the model to learn a good decision boundary. By considering the node type of our negative samples, we dramatically reduce the sample space (i.e., sampling negative t' only from the same node type of t). This reduction ensures that the negative samples remain challenging enough to drive further improvements in the model's performance. Another adaptation we made to better suit the biomedical KG reasoning scenario is to include a BRG for predicting links of interest. Any background knowledge can be used here, such as a PPI detailing the regulatory relationships between proteins. Prediction can be made without, or more knowledge can be supplied in form of a BRG which is only used for message passing (Figure 2.15C). For example, in predicting a link between nodes h and t, the graph between type 1 and 2 nodes can be used (without BRG), yielding one path such as given in Figure 2.15D. Otherwise, a BRG is included making use of a graph that comprises relations between type node 2 and 3 (Figure 2.15E). The tool can be found at `https://github.com/emyyue/BioPathNet`.

2.7.2 Model evaluation metrics

We used KGC metrics to measure the performance of BioKGC and all other models for comparison. These typically include MR (Equation 2.33), MRR (Equation 2.34), and Hits@k (Equation 2.35), as detailed in the background Section (Section 2.4.6.3). These metrics are ranking-based

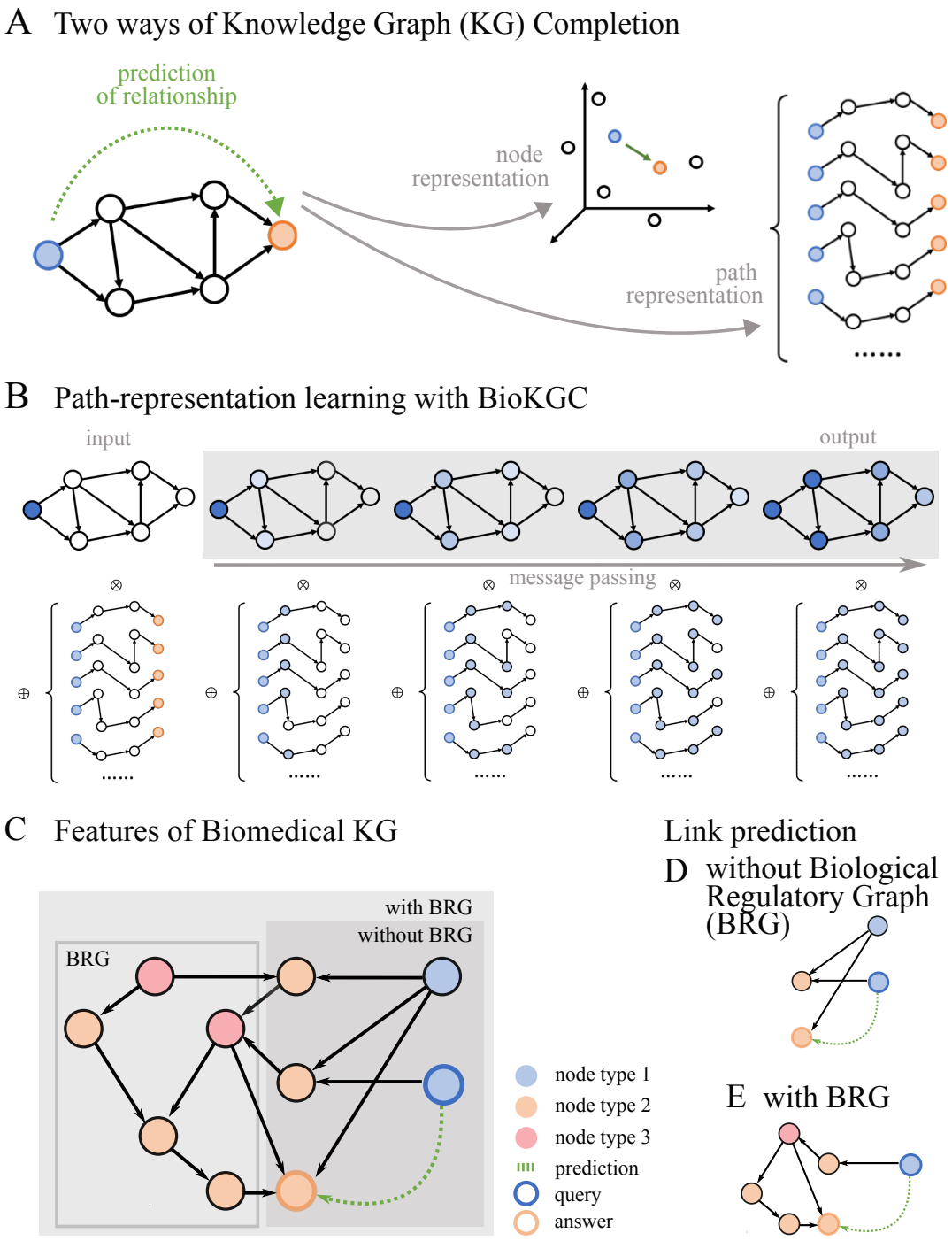


Figure 2.15: **Link prediction in biological KGs.** A) KGC can be performed by learning node and path representations. B) As a path-based method, BioKGC is a single source node-dependent framework that computes pairwise relationships from the source node to all target nodes using message passing. The computed output for each node is the path representation from the source to the target node. These are parsed to an MLP to learn positive and negative relationships between nodes in a supervised manner (not shown). C) As an extension of NBFNet, the BioKGC model developed explicitly for biomedical KGs considers two edge types: training and message-passing. A BRG with more knowledge can be leveraged for message passing on top of the training edges. Furthermore, node types can be defined for improved negative sampling. Examples of paths used for prediction between query and answer nodes (D) without and (E) with a BRG. (Adapted from and curtesy of Zhaocheng Zhu and adapted from Hu et al. [2024])

and quantify the method's ability to rank positive triplets among all candidate triplets or, in the filtered ranking protocol, among all negative triplets [Bordes et al., 2013]. The metrics Hits@k and Recall@k provide the proportion of ground truth positive triplets in the top k predictions.

The MR is calculated by taking the average of the ranks of all positive triplets and, therefore, treats all ranks linearly. The MRR ranges between $[0, 1]$ and is the average of the reciprocal ranks, which is 1 divided by the ranks. The reciprocal rank decreases non-linearly as the positive triplets move down the ranking list. Thus, greater emphasis is placed on the correctness of the top predictions, and a higher MRR means that a model can consistently rank positive triplets higher, even though some triplets may be ranked poorly. Conversely, if the MR of positive triplets is poor, the general model might perform well, but the metric is skewed by triplets with very poor rankings. Models are usually evaluated and selected based on the best MRR, allowing the model to prioritize ranking many triplets in the top predictions rather than improving some triplets with poor rankings.

The other evaluation metrics used were not specific to the KGC field based on ranking but rather commonly known metrics in ML and DL. Commonly in KGC the conditional probability is modeled, i.e. the probability of the tail entity *given* the head entity h and the relation r. However, there are cases where it is reasonable to evaluate the the joint probability p(h, t, and r). To ensure consistency with the model we compared against, we employed AUPRC to summarize precision and recall across various probability thresholds. Additionally, we evaluated specificity and F1 score at a fixed probability threshold of 0.5 in the second part of the analysis (Section 2.2.6.2).

2.7.3 Visualization

Compared to embedding-based methods, one advantage of BioKGC as a path representation learning method is its ability to visualize the most important paths for each prediction. For the paths between a given query and its predicted answer entity, we use the prediction gradient to assign an importance score to each path. The local landscape of the model is approximated with a linear model over the set of paths, with the weights computed by the partial derivative of the path prediction. Concretely, path importances is estimated as the sum of the edge importances derived via auto-differentiation [Zhu et al., 2021].

$$P_1, P_2, ..., P_k = top - k_{P \in P_{uv}} \frac{\partial p(u, q, v)}{\partial P} \tag{2.41}$$

In the visualization plot, we selected the top 10 most important paths ranked by gradients, with edge width representing the frequency of edge appearances across paths. Additionally, highlighted in red is the path with the highest weight, and nodes are colored by their node type. This level of interpretability allows predictions to be evaluated for their biological plausibility, facilitating hypothesis generation or validation in laboratory experiments.

2.7.4 Case studies: Functional annotation and drug repurposing

To demonstrate the capabilities of BioKGC to perform link prediction biomedical KGs, we applied our method to two different tasks. While the first task of functional annotation presents a proof-of-concept, zero-shot drug repurposing marks a more difficult real-life application scenario.

2.7.4.1 Functional annotation

As a proof-of-concept that KGC methods are useful for predicting links in biomedical KGs based on the already present information, we applied BioKGC to the functional annotation task, retrieving connections between cellular pathways and genes. For this task the KG we used consisted of a BRG and the functional annotations themselves. First, we obtained the BRG detailing the regulatory relationships between genes and chemicals from Pathway Commons [Cerami et al., 2010, Demir et al., 2010, Rodchenkov et al., 2020] in a simple interaction format given as triplets `(node1, relationType, node2)` (PathwayCommons12.All.hgnc.sif.gz) from www.pathwaycommons.org. Second, we downloaded the KEGG functional annotations (CPDB_pathways_genes.tab) from ConsensusPathDB (http://cpdb.molgen.mpg.de/) working with gene symbols (HUGO Gene Nomenclature Committee) [Kamburov et al., 2009, 2011].

During data preprocessing, underrepresented KEGG pathways that denote fewer than 10 annotations per gene were removed. Of the combined graph of BRG and KGG, we retained only the largest connected component, which led to the removal of 11 nodes that were part of smaller components containing only 2–3 nodes. Lastly, triplets were excluded if they contained genes present in the validation or test set but not available as supervision and message passing nodes during training. Our model, BioKGC, was trained on 70% of the P-G triplets. Validation was performed on 10% and testing on 20% of the P-G triplets. Two scenarios were considered, once making use of the underlying BRG and once without as additional message-passing graph. When the BRG was not used, triplets containing genes present in the validation or test but not the supervision training set were excluded.

We further benchmarked BioKGC against three KGE methods (TransE, DistMult, and RotatE, Section 2.4.5) and R-GCN (Section 2.4.6.2, which uses a GCN for message passing to learn node embeddings. To investigate the robustness, we ran each method with five different seeds and reported the mean ± standard deviation. We visualized the most important paths as a local subgraph to understand how the predictions were made. To exemplify how BioKGC works for this application, Figure 3.10A shows that we can use the path *function*→ *interacts with*→*phosphorylates* over the nodes a and b to predict that the pathway h is likely to be a functional annotation of the gene c.

2.7.4.2 Drug-disease prediction

After conducting a proof-of-concept and completing the KEGG knowledge base, we turned to an especially challenging link prediction scenario. In the second part of our study, we evaluated BioKGC's performance against TxGNN, a state-of-the-art model for drug repurposing in the zero-shot prediction setting. The PrimeKG database was used, which was assembled from

various sources, combining databases on genes, gene function, diseases, drugs, and phenotypes [Chandak et al., 2023] (Figure 3.11A). training, validation, and testing sets were created using TxGNN's native code that mimics a zero-shot prediction setting. Triplets containing the relation types indication, contraindication, and off-label use for a certain disease area were removed from training, alongside 95% of remaining connections to phenotypes or proteins [Huang et al., 2023]. The authors deviced these splits to mimic diseases with little molecular characterization and no known treatments. We used five zero-shot disease areas of adrenal gland, anemia, cardiovascular, cell proliferation, and mental health.

Next, TxGNN constructs the reverse edges, making the graph undirected. However, we excluded the reverse relations, as BioKGC inherently generates reverse triplets during the reasoning process. Furthermore, TxGNN consists of a pre-training and a fine-tuning phase. During the former, all triplets are used for learning. During the latter, only triplets containing the drug-disease relations indication, contraindications, and off-label use are used. BioKGC does not divide in these two phases, instead non-drug-disease relations (i.e. BRG) are used for message passing and drug-disease triplets are considered for supervision. We considered following hyperparameters searches for our model based on the performance (i.e. MRR) on the validation set: aggregator function $= \{sum, pna\}$, adversarial temperature $= \{0.5, 1, 2, 5\}$, parameters dependent $= \{yes, no\}$, and number of negative samples $= \{32, 64, 128\}$ and number of hidden layers $= \{2, 4, 6, 8\}$. Following the example of TxGNN and ensuring robustness, five seeds in data split were used. Results were reported as mean $\pm$ standard deviation of performance metrics.

We evaluated our models against TxGNN regarding AUPRC, specificity, F1 score, and Recall@k, which are native to TxGNN's code base. TxGNN computes AUPRC using two distinct strategies. The first, referred to as `AUPRC 1:1`, compares each positive ground truth item against one negative sampled from the list of drugs within a disease area. This metric captures the model's ability to differentiate a positive item from a randomly selected negative in a direct comparison. The second strategy, referred to simply as `AUPRC`, considers all positive and negative ground truth items across the dataset. Unlike `AUPRC 1:1`, this approach evaluates the model's ability to distinguish positives and negatives in a comprehensive manner, accounting for class imbalance providing a balanced assessment of recall and precision. We predominantly used the latter, as it better reflects real-world scenarios where identifying therapeutic opportunities from a set of drugs is critical.

After a quantitative evaluation, we investigated the top therapie options for acute lymphoblastic leukemia (ALL), gastric cancer, and lymphoma within the cell proliferation split and examined known and unknown predictions. We discovered that the data split obtained over TxGNN only exhibits a near-zero shot setting, as treatments for some related diseases remain. Refining the inference graph in a small experiment, we excluded triplets of related diseases by removing disease whose node name could be matched by searching the string *tumor, lymphosarcoma, neoplasm, teratoma, leukemia, cancer, cytoma, lymphoma*, or *carcinoma*. This way we examined the paths the model relied on when disease similarity is excluded from predictions, with the goal of uncovering novel biological insights. Next, we illustrated the open-world assumption in biomedical KGs, calculating the MRR@k (MRR at top k predictions) for lymphoma before and after considering treatments for specific lymphoma subtypes, such as ALL, Hodgkin's lymphoma, and primary central nervous system lymphoma.

We additionally created a custom data split using TxGNN's `disease_eval` code to assess perfor-

mance in predicting drugs for the neurodegenerative disorder Alzheimer's disease (AD). Given the devastating nature of AD and the limited understanding of its causes and treatment options, we evaluated the model's predictions with expert input to better understand the strengths and limitations of BioKGC.

Chapter 3

Results

In this chapter, I will detail the results from the three methods and application to biological problems. First, the focus will be on identifying key players in the complex disease MDD using KiMONo (Section 1.4 and 2.5). This section is based on and partly identical to Ogris et al. [2021]. Second, after the integration of multi-omic data, we will explore the embedding of nodes from multi-modal network to efficiently contextualize COVID-19 related genes (Section 1.5 2.6). This section is based on and partly identical to Hu et al. [2022]. Third, coming to the task of link prediction, we will model the specific relationships between nodes explicitly to learn from existing knowledge to predict missing but true links in the functional annotation and drug repurposing task (Section 1.6 and 2.7). This section is based on and partly identical to Hu et al. [2024].

3.1 Analysis of multi-omic network identified key players in the complex disease MDD

To obtain an integrated view of the complex disease MDD, we used the data from a patient cohort with controls to establish a multi-omic network (Section 2.5).

3.1.1 MDD patient cohort derived multi-omic network

We applied KiMONo to a cohort of 107 individuals, comprising both healthy controls and patients. This dataset featured 4,247,909 imputed SNPs, 12,418 transcripts, and 320,481 methylation sites (after excluding those with the lowest variance [i.e., the lowest 25%]). We also incorporated clinical factors, including BMI, age, sex, diagnostic status, and cell type composition, into our network inference process.

Ensuring the quality of the selected features, we applied filters for the coefficients β within the range of -0.02 to 0.02 and R^2 values less than 0.1. Consequently, the MDD network comprised $9,943$ gene models, with a median R^2 value of 0.184. Notably, a few models exhibited very high R^2 values exceeding 0.75. Within this network, we identified $7,837$ methylation sites and $3,749$ SNPs as first-order links, and $5,336$ gene transcripts and $4,351$ methylation sites as second-order

links. Additionally, all biological covariates were observed throughout the entire network (Figure 3.1A, B).

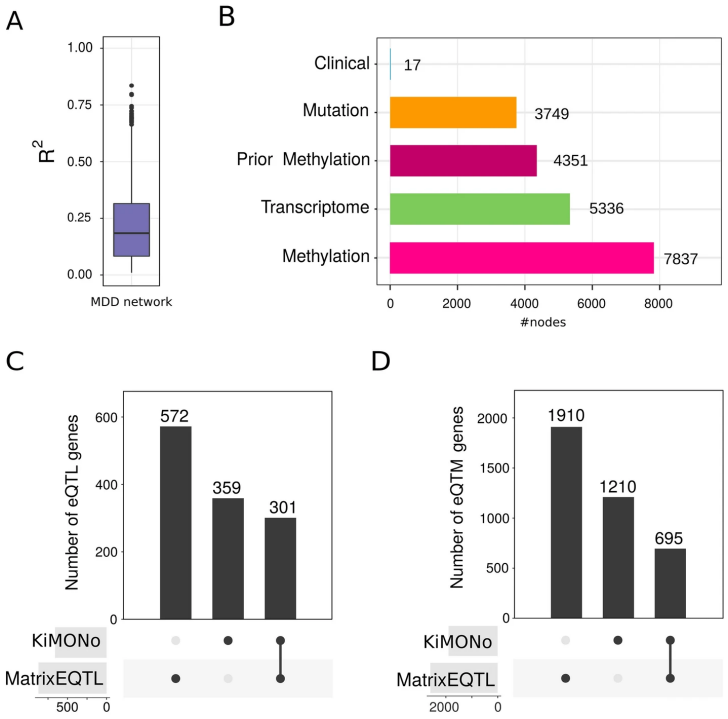


Figure 3.1: **Characterization of the major depressive disorder (MDD) multi-omic network and comparison to two-omic integration method:** A) The performance of the inferred disease network on MDD of all gene models ($n = 9943$) after filtering for $-0.02 < \beta < 0.02$ and $R^2 < 0.1$. B) The features retained after regularization derived from the omic levels comprising first-order links (methylation sites and SNPs) and second-order links (methylation sites, SNPs, gene expression, and clinical features). C) The numbers of eQTL and eQTM genes identified by matrixEQLT and KiMONo. (From Ogris et al. [2021])

3.1.2 Information gain of multi-variate models compared to two-level omic integration technique eQTL and eQTM

We used pairwise models to identify eQTL and eQTM genes to benchmark against state-of-the-art techniques. We then contrasted these findings with those of the KiMONo approach. Using identical proximity constraints for MatrixEQTL and KiMONo, we identified 873 and 660 eQTL genes, respectively, with an overlap of 301 genes (Figure 3.1C). Furthermore, we identified 695 eQTM genes shared between the two methods, with 1910 unique genes identified by MatrixEQTL, surpassing the 1210 identified by KiMONo (Figure 3.1D). Notably, most overlapping genes and those exclusively identified by KiMONo were modeled using multivariate models, incorporating information not only from genetic predisposition but also other omics layers such as methylation, SNPs, and gene expression.

3.1.3 Top most important genes capture MDD disease genes

The top 20 genes, ranked by the importance measure, betweenness, performed better than the average model. Their R^2 values ranged between 0.202 and 0.798, with the median at 0.525. In contrast, the average R^2 across all models was 0.539 (Figure 3.2A). The features the regularized regression models selected encompassed information from various omic-data levels, including methylation sites, SNPs, clinical features and gene expression. Furthermore, the top 20 hits consistently featured methylation sites with long-distance effects, gene expression linked via indirect connections, and biological information (Figure 3.2B).

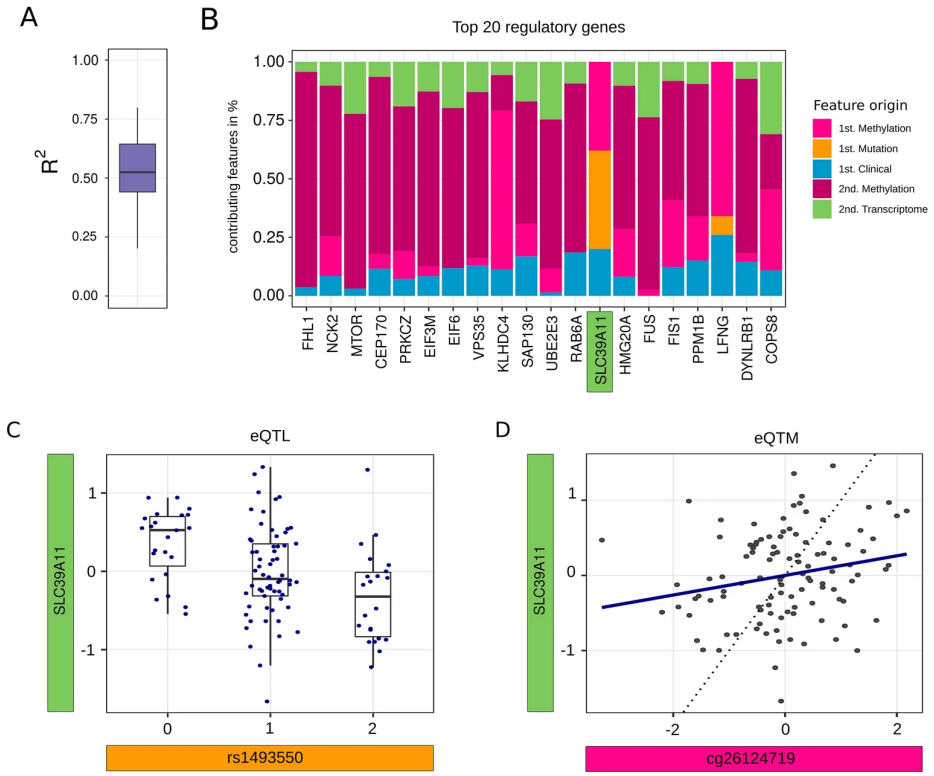


Figure 3.2: **Top 20 most important nodes and putative multi-omic interplay with SLC39A11:** A) The performance of the top 20 models, ranked by the betweenness centrality measure, on MDD after filtering for $-0.02 < \beta < 0.02$ and $R^2 < 0.1$. B) The composition of retained features (in %) after regularization derived from the different omic levels. The putative interplay between gene expression and (C) SNPs and (D) methylation sites identified with KiMONo; the dotted line represents a correlation of 1. (From Ogris et al. [2021])

The effectiveness of our approach became apparent when we examined the connections uniquely retrieved by KiMONo rather than by the pairwise models of MatrixEQTL. When residual effects were removed across all other features in the multi-omic models, we could clearly show connections between the expression of solute carrier family 39 member 11 (*SLC39A11*) on chromosome 17 and the SNP rs1493550 and the methylation site cg26124719, both of which are located within one of its introns (Figure 3.2C and D).

Half of the top 10 hits had previously been associated with depression or pathways involved in the pathogenesis of MDD or related bipolar disorder (BP) (Table 3.1). The most significantly enriched Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway was endocytosis, with an FDR value of $4.832e^{-8}$, which is crucial in synaptic plasticity and has been identified as a key

Table 3.1: The top 10 genes ranked by their betweenness, as a measure of importance and influence in the condition-specific multi-omic network generated by KiMONo, and evidence for their involvement in MDD and BP. (From Ogris et al. [2021])

Rank	Gene symbol	Gene name	Location	MDD	BP	References
1	FHL1	Four and a half LIM domains 1	chrX:136,146,702–136,211,359	-	-	
2	NCK2	NCK adaptor protein 2	chr2:105,744,912–105,894,274	-	-	
3	MTOR	Mechanistic target of Rapamycin kinase	chr1:11,106,531–11,262,557	Yes	-	[Abelaira et al., 2014, Ignácio, 2016]
4	CEP170	Centrosomal protein 170	chr1:243,124,428–243,255,406	-	-	
5	PRKCZ	Protein kinase C zeta	chr1:2,050,411–2,185,395	-	Yes	[Hapak et al., 2019, Kandaswamy et al., 2012]
6	EIF3M	Eukaryotic translation initiation factor 3 subunit M	chr11:32,583,767–32,606,2	Yes	-	[Terracciano, 2010, Varinthra and Liu, 2019]
7	EIF6	Eukaryotic translation initiation factor 6	chr20:35,278,906–35,284,985	-	-	
8	VPS35	Vacuolar protein sorting-associated protein 35	chr16:46,656,132–46,689,518	Yes	-	[Wang, 2012]
9	SAP130	Spliceosome-associated protein 130	chr2:127,941,217–128,028,120	Maybe	-	[Relja et al., 2018]
10	KLHDC4	Kelch domain containing 4	chr16:87,696,485–87,765,997	Yes	-	[Relja et al., 2018, Roberson-Nay et al., 2018]

component in developing stress-related disorders such as MDD [Duman et al., 2016, Hua, 2013]. The second most important pathway was autophagy ($FDR = 2.606e^{-6}$), which is essential for the central nervous system. One study has indicated the influence of antidepressant treatments on autophagy [Gassen and Rein, 2019]. Additionally, axon guidance ($FDR = 1.054e^{-3}$) was identified as a significant risk factor for MDD. Stress can impact brain structure and function, contributing to this connection [Breen, 2018, Engle, 2010].

A

target	predictor	relation	mean_value	mean_rsq
RPL6	RPS9	prior_expr	0.0004894930	0.9127852
RPL6	RPS3	prior_expr	-0.0113043323	0.9127852
RPL6	RPS11	prior_expr	-0.0002571735	0.9127852
RPL6	RPL7	prior_expr	0.0129995783	0.9127852
RPL6	ILF2	prior_expr	-0.0004909220	0.9127852
RPL6	HNRNPU	prior_expr	0.0006112889	0.9127852
RPL6	FGF2	prior_expr	-0.0040520966	0.9127852
RPL6	Pituitary	expr_tissue	0.0182454607	0.9127852
RPL6	Muscle_Skeletal	expr_tissue	0.0002804242	0.9127852
RPL6	Lung	expr_tissue	0.0018316824	0.9127852
RPL6	PRS_pneumonia	expr_prs	0.0351627872	0.9127852
RPL6	PRS_drinking	expr_prs	0.0324385794	0.9127852
RPL6	PRS_diabetes2	expr_prs	0.0238945022	0.9127852
RPL6	PRS_asthma	expr_prs	0.0122384741	0.9127852
RPL6	PRS_COPD	expr_prs	0.0509888007	0.9127852
RPL6	diabetes2	expr_como	0.0042451691	0.9127852
RPL6	diabetes1	expr_como	0.0711155421	0.9127852
RPL6	PRS_pneumonia	expr_como	-0.0052523170	0.9127852
RPL6	age	expr_bio	-0.0008140422	0.9127852

B

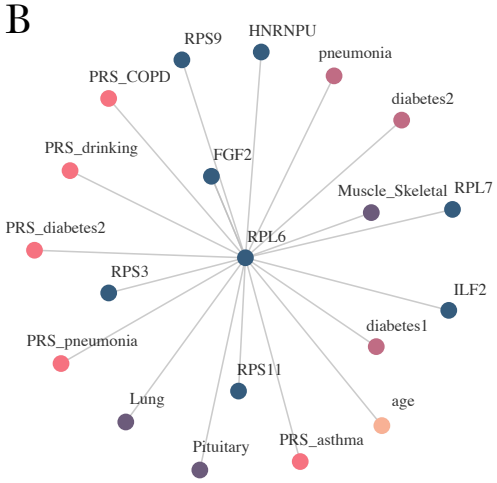


Figure 3.3: **Example of a KiMONo's gene model** selected features over 30 runs of stability selection: RPL6, with a R^2 of 0.912 and the beta coefficient of features, given as output table from sparse group lasso and B) the statistical associations displayed as edges between RPL6 and its explanatory variables from different modalities of genes, phenotypes, tissues, disease states and PRS. (From Hu et al. [2022])

3.2 Embedding multi-modal network embedding contextualized COVID-19 genes

After the comprehensive view of disease pathology of MDD using multi-omic integration methods for the identification of most important genes, we turn to the disease COVID-19 and analyze the relationships between entities over multi-omic embeddings. COVID-19 affects multiple organs and causes symptoms related to diverse tissues, such as liver, lung, hematological neurological, heart and kidney diseases. To gain insights into these multi-modal characteristics, we used the prepandemic GTEx cohort, which comprises nearly 1,000 individuals and encompasses various disease diagnoses, referred to as comorbidities in the context of SARS-CoV-2 infection. We applied our novel ML framework that constructs a multi-modal network and embeds the nodes into a low-dimensional space, evaluated it in its biological plausibility and revealed the multi-modal context of COVID-19 genes (Section 2.6).

3.2.1 Disease state and PRS capture differential information

Using our novel framework, we successfully integrated data from various modalities, including phenotypes, gene expression profiles from 43 tissues, genetic risks, and existing disease diagnoses. We calculated PRSs for individuals in the GTEx dataset with available genotypes using a genome-wide scoring approach to capture the genetic risk associated with disease development. We computed PRSs for 24 GWAS studies on diseases and traits that are known or suspected to be associated with an increased risk of severe COVID-19. PRSs were also calculated based on three COVID-19 GWAS studies.

We analyzed correlations to explore the relationships between disease states, PRSs, and phenotypes. High correlations existed between the PRSs of some diseases, such as schizophrenia with

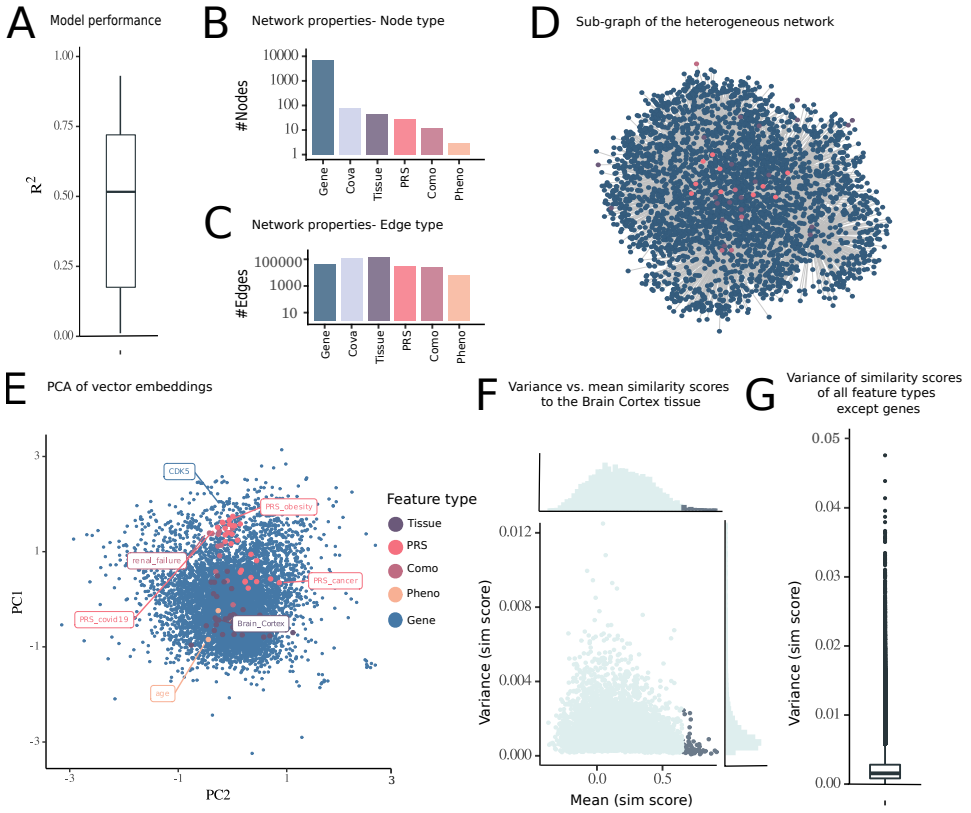


Figure 3.4: **Characterization of the GTEx multi-modal network:** Embedding of the multi-modal network of 13,486 samples from 793 individuals in the GTEx population cohort, consisting of the data modalities gene expression and PRS, COVID-19 comorbidities, phenotypes, and tissues. A) The performance of sparse-group LASSO models from KiMONo, expressed as R^2 (the variance explained). The number of (B) nodes and (C) edges of the inferred network. D) The subnetwork of the complete multi-modal network. E) Full network embedding (one representative run out of 100) shown as a PCA plot with obesity, COVID-19, and cancer (PRS), *CDK5* (gene), brain cortex (tissue), and age (phenotype) highlighted. F) Similarity scores for one node of interest (brain cortex) against all others plotted across 100 runs. Nodes with high similarity exhibit low variance, as given by the marginal density plot to the right. G) Variance of all non-gene nodes similarity. (From Hu et al. [2022])

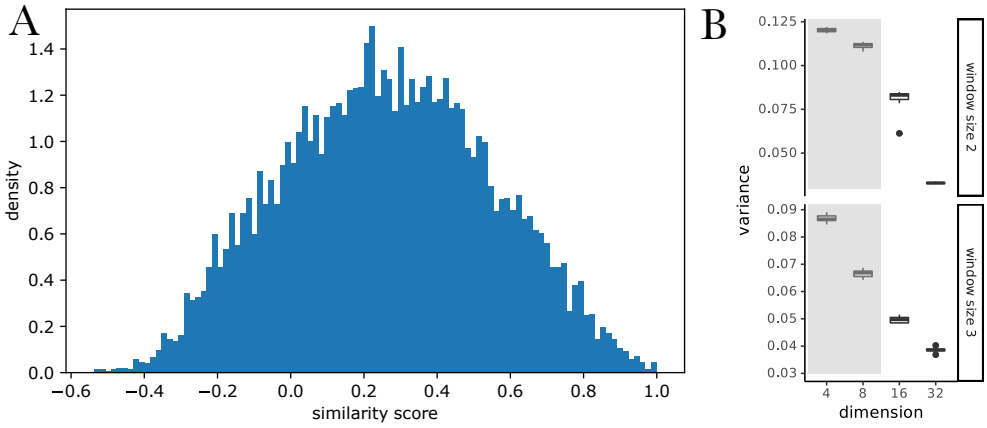


Figure 3.5: **Hyperparameter search for network embedding over grid search, on a smaller brain network:** A) Distribution of similarity scores between 10,000 random node pairs (window size = 2, embedding dimensionality = 16), variance = 0.080. B) Variance of 10,000 random node pairs across different hyperparameter settings: window size = [2,3] and dimensionality of embedding = [4, 8,16, 32] over 10 embedding repetitions. Gray overlay indicates deviation from normal distribution of similarity scores and thus, discarded for hyperparameter search. (From Hu et al. [2022])

BP (0.97), MDD (0.62), and coronary artery disease (0.53). The PRSs were also correlated for COVID-19 severity (0.76), T2D with type I diabetes (T1D) (0.38), hypertension (0.46) and stroke (0.24). However, weak correlations existed between actual disease states (as provided by GTEx) and genetic risk (PRS). For example, the correlations between T2D and diseases such as renal failure, hypertension, and clinical factor age were 0.22, 0.35, and 0.25, respectively. Similarly, the correlations between T1D, asthma, and hypertension with their respective actual disease states were also weak (0.19, 0.09, and 0.14, respectively). Indeed, the weak correlation between genetic risk (as measured by PRS) and disease status highlights the importance of integrating complementary information into disease studies. While the PRS captures the genetic risk for developing a disease, the real-world development is also influenced by external environmental factors.

The multi-modal GTEx data was integrated using KiMONo's inference algorithm. Figure 3.3 provides an example of a single gene model. Models were filtered based on performance, β values, and stability selection. Out of all 7,251 gene models, 7,236 with 461,216 edges were retained after filtering. Reverse models were run, and statistically significant associations were retained. The median R^2 value for all models was 0.52 (Figure 3.4A). The resulting network consisted of phenotypes ($n = 3$), genes ($n = 7,202$), covariates ($n = 77$), diseases ($n = 12$), tissues ($n = 43$) and PRS ($n = 27$) (Figure 3.4B). Genes were the most common node type in the network, and the most common edge types were between genes and tissue variables ($n = 62,902$), genes and covariates ($n = 44,965$), genes and other genes ($n = 18,521$), genes and PRS ($n = 11,613$), genes and disease states ($n = 10,553$), and genes and phenotypes ($n = 2,817$) (Figure 3.4C).

To determine the best parameters for the embedding process in the multi-modal network, we conducted a grid search on a smaller brain network. The optimal parameters for the models were a window size w of 2 for defining positive examples and an embedding dimensionality d of 16. We evaluated the performance of these parameters by measuring the normally distributed variance, which exhibited the highest value (mean = 0.080) across 10 repetitions (Figure 3.5A). This finding reflected that the selected parameters captured the maximum information content

without overfitting the data. To account for the variability in stochastic random walks, we performed network embedding 100 times using these optimal parameters in the multi-modal network inference with KiMONo.

To visualize the 16-dimensional embedding space in a simplified manner, we chose one random run out of 100 and subjected it to a PCA. The nodes were then plotted in the two-dimensional space defined by the first and second PCs. The combined variance explained by these two components was only 20.17%, as the number of dimensions d was optimized for in hyperparameter search (Figure 3.4D). The genes were spread throughout the PCA plot, well embedded among other modalities. Some diseases and PRS showed high values along PC1, including the PRSs for COVID-19, obesity, and renal failure, whereas the PRSs for cancer and several tissues had lower PC1 values (Figure 3.4E). It is important to note that while linear PCA cannot fully capture the nonlinear nature of the embedding space, it provides an approximation of the space. Additionally, we computed the maximum similarity score between the node brain cortex tissue and all others across the 100 runs. We observed that the topmost similar nodes are robust with little variance across runs (similarity score > 0.65) (Figure 3.4F). Furthermore, low variances in similarity scores between non-gene and all other nodes showed another aspect of robustness of the embedding space (Figure 3.4G).

3.2.2 Evaluation of embedding space

In order to verify the plausibility of the embedding space, upon which we base the further analysis of the COVID-19 related genes and their multi-modal context (Section 3.2.3), we first conducted three checks. The first is based on the ability of the embedding space to recapitulate tissue specific genes in the proximity of tissue nodes (Section 3.2.2.1). The second check verifies the context around disease states (Section 3.2.2.2) and thirdly, we investigated the PRS associated genes (Section 3.2.2.3).

3.2.2.1 Evaluation based on tissue specific genes

To validate the obtained embedding space of our multi-modal network, we analyzed the extent to which similarity scores between gene nodes and tissue nodes reflected established tissue-specific gene expression patterns. We accomplished this by comparing the similarity scores of genes with the brain tissue node and identifying the most and least similar genes. We then compared these sets of genes with those previously reported to exhibit enhanced expression in brain tissues in The Human Protein Atlas database. The set of most similar genes had a higher frequency of genes known to be functionally relevant in the brain than compared to the set of least similar genes. The odds ratio of enrichment was 4.33-fold higher within the most similar nodes than within the least similar nodes when considering the top and bottom 200 nodes ($mtop = 20$, $mbottom = 5$, chi-square test $p-value = 0.0038$), and reached 12.24-fold when considering the top and bottom 100 nodes ($mtop = 11$, $mbottom = 1$, chi-square test $p - value = 0.0074$) (Figure 3.6A). This trend was also observed for other tissues, such as the liver, which had many samples in the GTEx dataset and a substantial number of genes with tissue-enhanced expression (Figure 3.6A). Furthermore, brain-specific genes was confirmed by their enhanced expression values within the GTEx dataset (Figure 3.6B). To illustrate the embedding space with the brain cortex tissue node

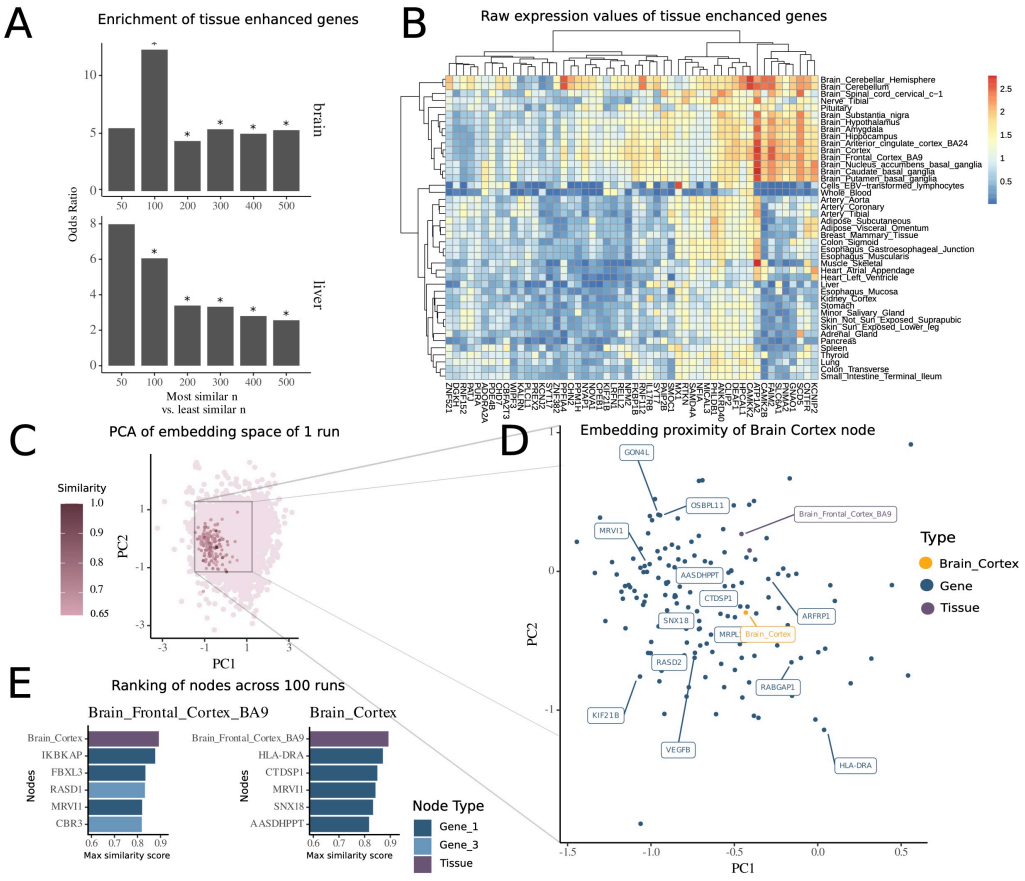


Figure 3.6: **Identification of tissue-associated genes in multi-modal network embedding:** A) Tissue-enhanced genes enrichment in the brain and liver within the top n most similar genes compared to the bottom n least similar genes, as measured by their mean cosine similarity score across 100 embedding runs. B) Higher median expression in the GTEx dataset of 50 genes within the top 500 most similar nodes in brain tissues. C) Embeddings from one representative run of all nodes of the multi-modal network with nodes with high similarity to the "brain cortex" node highlighted. D) Zoom in on the embedding space around the "brain cortex" node with the 15 top most similar nodes highlighted. E) The most similar nodes ranking for "brain frontal cortex BA9" and "brain cortex" summarized over 100 embedding runs. 1-hop neighbors are highlighted in dark blue, and 3-hop neighbors in light blue. (From Hu et al. [2022])

as center, we visualized the entire embedding from a single run using the first and second PCs of a PCA (Figure 3.6C) and then concentrate only on most similar nodes in the local embedding space (Figure 3.6D).

Here, we found the major histocompatibility complex, class II, DR alpha (*HLA-DRA*) gene in close proximity to the brain cortex. As indicated by a prior study [Fan et al., 2017b], the gene exhibited elevated expression in gliomas. Similarly, the brain frontal cortex BA9 node was among the most similar. To illustrate the application of similarity scores in contextualizing genes and other factors, we extracted the top nodes for both brain cortex and brain frontal cortex from 100 embedding runs, ranked based on maximum similarity score (Figure 3.6E). Previous studies have annotated the top genes in both brain regions associated with mental diseases or involved in brain functionality. Notably, the identified nodes included Ras-related dexamethasone-induced 1 (*RASD1*), encoding a GTPase which is overexpressed as a protein in the frontal cortex [Fishilevich et al., 2016] and associated with the nitric oxide synthase 1 (*NOS1/nNOS*) signaling pathway at neuronal synapses. Interestingly, despite being a 3-hop neighbor in the KiMONo multi-modal network rather than a direct neighbor, *RASD1* exhibited high similarity to the brain frontal cortex node (similarity score = 0.84). This observation highlights the network embedding's ability to capture relevant, non-linear relationships between nodes, even when relatively distant in the underlying network.

3.2.2.2 Evaluation based on context of known diseases - comorbidities

We extend to illustrate evaluation of the embedding space by delving deeper into the associations to nodes related to selected diseases and PRSs. For example, ischemic heart disease exhibited the closest proximity to hypertension, COPD, and cerebrovascular disease (Figure 3.7A). Among the genes and neighboring nodes within 1-hop and 2-hop distances in the underlying network, we identified minichromosome maintenance complex component 5 (*MCM5*), chromatin licensing and DNA replication factor 1 (*CDT1*), and cAMP-responsive element-binding protein 1 (*CREB1*). These genes have been previously associated with ischemic heart disease. Notably, CREB1 has been recognized as a robust genetic predictor for heart rate response, functioning during cardiac memory and contraction [Haidar et al., 2020, Rankinen et al., 2010, Urbanek et al., 2005]. Moreover, among the top 15 most similar nodes, we discovered nodes related to heart failure (genetic predisposition, i.e. PRS) and predisposition for COVID-19 hospitalization.

Interestingly, the disease node representing MDD was located amidst gene nodes in the PCA, and no other node modalities appeared in the top 15 most similar nodes (Figure S3.7B). The genes with the highest similarity to MDD included mammalian STE20-like kinase-1 (*MST1*), a kinase that promotes apoptosis, and aryl hydrocarbon receptor nuclear translocator like 2 (*ARNTL2*), which is involved in circadian clock regulation and associated with suicide risk in MDD via three GWAS. Additionally, among the most similar genes was cytochrome P450 family 2 subfamily E member 1 (*CYP2E1*) being an important protein in the microsomal oxidation system. Mutual pathomechanisms in MDD and nonalcoholic fatty liver disease was evaluated by Shao et al. [2021], highlighting their role in mediating and promoting each other's progression.

As another example, T2D exhibited high similarities to T1D, pneumonia, and the tissue tibial artery (Figure 3.7C). Genes within 1- to 3-hop distances from the T2D node have been previously described in the literature and associated with T2D. They include fibroblast growth factor

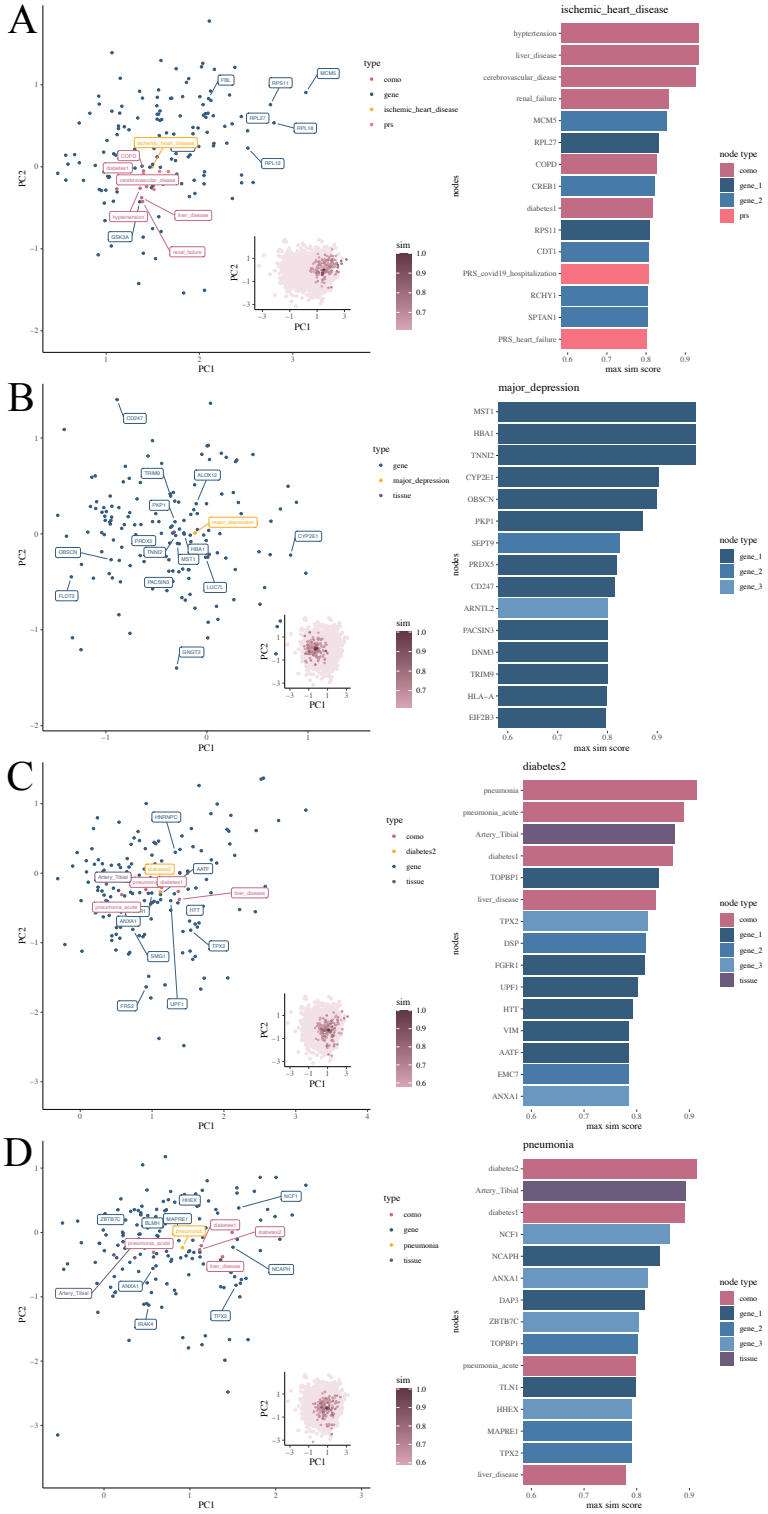


Figure 3.7: **Visualization of embedding space and top similar nodes of comorbidities** A) ischemic heart disease, B) major depression, C) Diabetes Type II and D) pneumonia. (From Hu et al. [2022])

receptor 1 (*FGFR1*), microtubule nucleation factor (*TPX2*), RNA helicase and ATPase (*UPF1*), and Huntingtin (*HTT*) [Hall et al., 2019, Li et al., 2020, Montojo et al., 2017, Tani et al., 2012].

Investigating a last example of disease states, we focus on pneumonia as a known pathology of COVID-19 (Figure 3.7D). Here, acute pneumonia, T1D, T2D, and the tibial artery were among the nodes with the highest similarities. It has been established that patients with diabetes have an increased risk of pneumonia [Ehrlich et al., 2009, Vardakas et al., 2007]. Furthermore, we found that neutrophil cytosolic factor 1 (*NCF1*), encoding a component of the NADPH oxidase complex, which further has been associated with fibrosis, inflammation, and pneumonia [Zamakhchari et al., 2016]. Additionally, annexin A1 (*ANXA1*), which functions in the innate and adaptive immune responses, has been identified as a regulator of the inflammatory response. Interestingly, *ANXA1* has also been proposed to be a prognostic biomarker for COVID-19, with its decrease observed in severe cases [Machado et al., 2020].

3.2.2.3 Evaluation based on known diseases - PRS

Next, we examined the nodes most closely associated with each PRS, which captures the genetic predisposition to a particular disease. This analysis aimed to identify the network elements most relevant to the genetic risk factors underlying our data. We observed that the top 15 nodes of each PRS often consisted of other PRS nodes. This inflated similarity could be attributed to the higher correlation among PRS nodes, as we had previously quantified before the embedding process (Figure 4.2). For example, when examining the node representing chronic kidney disease, 12 of the 15 most similar nodes were PRS for diseases including from Crohn's disease, coronary artery disease and COPD (Figure S3.8A). These diseases are known to be comorbidities of COVID-19 [Cai et al., 2013, Chen and Liao, 2016, Demir et al., 2014].

For the chronic kidney disease node the closest nodes were the genes scavenger receptor cysteine-rich type 1 protein M130 (*CD163*), glutathione S-transferase Mu 3 (*GSTM3*), and ribosomal protein S10 (*RPS10*). They were located within 1-hop, 1-hop, and 2-hop distances respectively. These genes have been identified as significant developmental factors of renal tissue carcinomas or have been recognized as biomarkers for inflammation [Mejia-Vilet et al., 2020, Serin et al., 2021, Tan et al., 2013].

For the cancer PRS, ornithine decarboxylase 1 (*ODC1*), identified as a 3-hop neighbor, emerged as the most similar gene. Previous research by Kim et al. [2017] suggested *ODC1* as a potential therapeutic target for endometrial cancer due to its frequent overexpression and its role in promoting cell proliferation. Additionally, other closely related nodes included the PRS for lung cancer, with a similarity score of 0.83 and alcohol abuse 0.79 (Figure 3.8B).

Regarding the schizophrenia PRS, it showed the highest similarity to the PRSs for MDD, BP, COVID-19, and obesity (Figure 3.8C). Indeed, the relationship between mental disorders and COVID-19 has gained attention, with studies reporting the highest odds ratios for susceptibility and mortality among individuals with severe mental disorders [Fond et al., 2021, Liu et al., 2021, Wang et al., 2021a]. This vulnerable group has been suggested to have lower immune function and worse resilience. This section has demonstrated how the similarity score can be used to contextualize disease states and PRS enabling the identification of disease-associated factors and genes

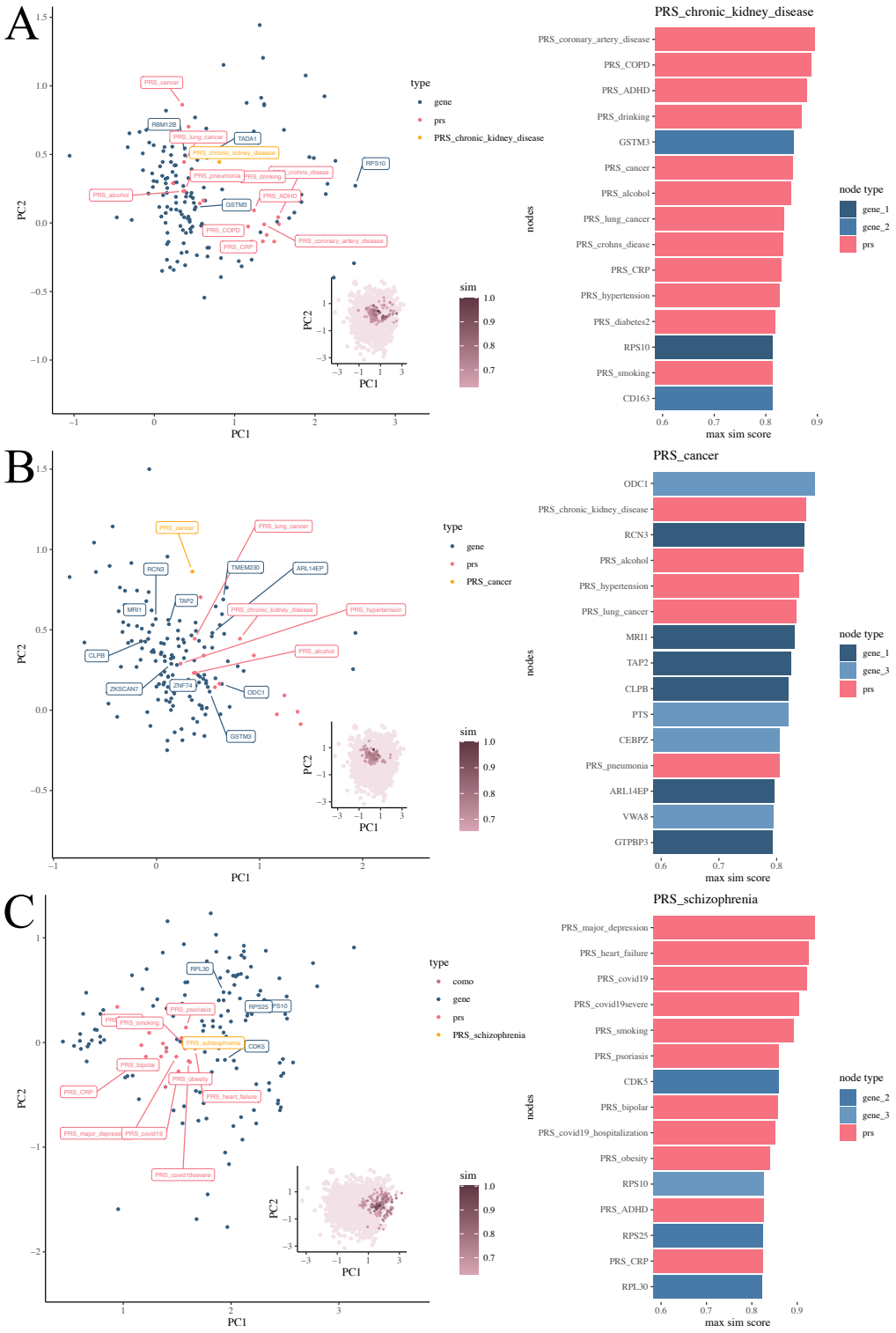


Figure 3.8: **Visualization of embedding space and top similar nodes of PRS** A) chronic kidney disease, B) cancer and C) schizophrenia. (From Hu et al. [2022])

3.2.3 Uncovering novel dependencies between COVID-19 and tissues, PRS, and COVID-19 comorbidities

After demonstrating the framework's ability to capture both tissue-specific and disease-specific genes and their associations, our next aim was to gain insights into host factors critical for

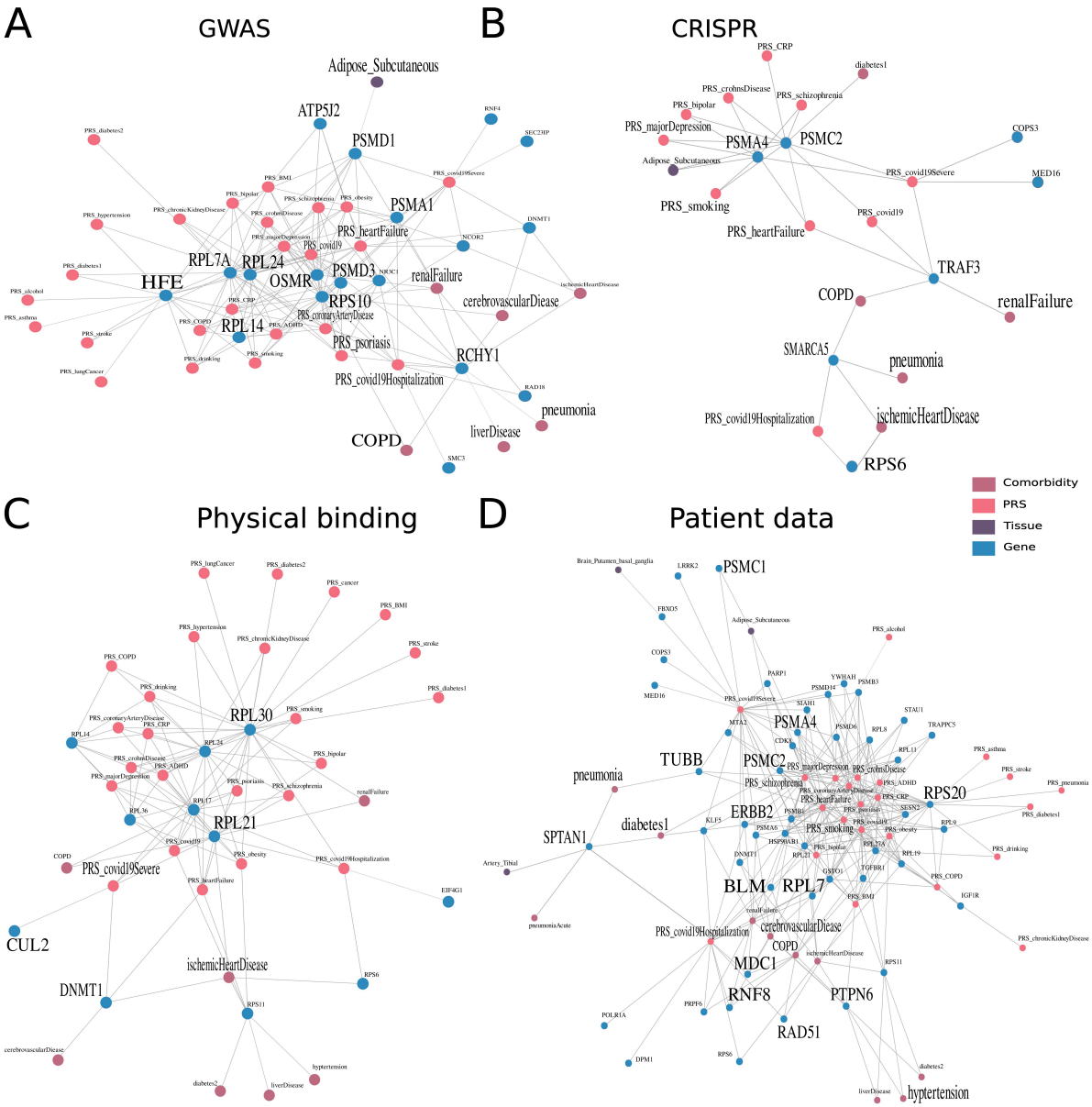


Figure 3.9: **Multi-modal context of established host factors crucial for COVID-19 derived from diverse public experimental and patient data sources:** Represented as a network, nodes in this context encompass tissues, PRSs, and COVID-19 comorbidities. An edge between two nodes was established when the similarity score within the embedding space exceeded 0.65 and demonstrated robust identification in 80 of 100 run repetitions. Similarity networks were constructed based on genes derived from (A) GWAS studies [The COVID-19 Host Genetics Initiative, 2020] and (B) CRISPR studies Schneider et al. [2021], Wei et al. [2021]. Additionally, (C) physical interaction experiments, including ribonucleoprotein capture and immunoprecipitation [Gordon et al., 2020, Lee et al., 2021], and (D) patient data from whole blood samples [Demichev, 2021, Di et al., 2020, D'Alessandro et al., 2020, Geyer et al., 2021, Messner et al., 2020, Overmyer et al., 2021, Shen et al., 2020, Wu et al., 2021] further contributed to the comprehensive understanding of this multi-modal context. (From Hu et al. [2022])

COVID-19 within the complex multi-modal context of the GTEx cohort data. To achieve this, we curated genes from various sources, including GWAS, CRISPR, physical interaction studies, and multiomics patient data. Then we explored the proximity of these gene sets in the embedding space. We represented the top associations of these genes with tissues, PRS, and diseases (considered COVID-19 comorbidities) as similarity graphs. The links were retained when two nodes passed the similarity score threshold of >0.65. We also emphasized the multi-modal context of genes linked to at least one of the three COVID-19 genetic predispositions: susceptibility, severity, or hospitalization.

First, we will detail the similarity network of genes obtained from GWAS studies (Figure 3.9A). Various ribosomal subunits and factors, including 60S ribosomal protein L7a (*RPL7A*), 40S ribosomal protein S10 (*RPS10*), 60S ribosomal protein L14 (*RPL14*), and 60S ribosomal protein L24 (*RPL24*) showed connections to numerous PRS. RPS10 was also linked to comorbidities such as cerebrovascular disease, COPD, and renal failure. Evidence suggests that viral NSP1 may block the ribosomal protein entry channel during infection, and thus prohibiting mRNA translation [Simeoni et al., 2021]. Another notable group of proteins consisted of proteasome components, including proteasome 26S subunit, non-ATPase 1 (*PSMD1*); proteasome 26S subunit, non-ATPase 3 (*PSMD3*); and proteasome 20S subunit alpha 1 (*PSMA1*).

In the embedding, *PSMA1* was closely associated to renal failure and adipose tissue and it has been reported to help maintain protein homeostasis via the ATP-dependent degradation of ubiquitinated proteins, including those of coronaviruses [Tiwari et al., 2021]. Another gene, Ring finger and CHY zinc finger domain containing 1 (*RCHY1*), was associated with different comorbidities, including cerebrovascular disease, COPD, renal failure, ischemic heart disease, pneumonia, and liver disease and with PRS for psoriasis, coronary artery disease and heart failure. The gene was derived from a GWAS study comparing hospitalized versus non-hospitalized COVID-19 cases. Similarly, our graph connected it to both the PRS of COVID-19 and COVID-19 hospitalization. RCHY1 is involved in the histone deacetylase 1 (HDAC1), E3-dependent ubiquitination and proteasomal degradation of proteins such as tumor protein 53 (TP53), and cyclin-dependent kinase inhibitor 1B (CDKN1B). It helps to regulate expression and cell cycle progression. While ribosomes and proteases were connected to many genetic risks, *RCHY1* exhibited similarities with various developed diseases known to be comorbidities of COVID-19.

The genes proteasome 26S subunit, ATPase 2 (*PSMC2*) and proteasome 20S subunit Alpha 4 (*PSMA4*) were associated with diverse PRS for diseases from heart failure to MDD (Figure 3.9B), which were identified by a CRISPR study [Schneider et al., 2021]. In addition, identified by Wei et al. [2021], tumor necrosis factor receptor-associated factor 3 (*TRAF3*), was associated with renal failure and COPD. Previous studies have implicated TRAF3 in signal transduction to activate immune and antiviral responses.

A member of the SWI/SNF family involved in ATP-dependent chromatin remodeling complexes (SWI/SNF-related, matrix-associated, actin-dependent regulator of chromatin, subfamily a, member 5 [*SMARCA5*]) and small ribosomal protein S6 (*RPS6*) were associated with hospitalization risk with COVID-19, as identified by Wei et al. [2021] and Schneider et al. [2021], respectively. These genes were also linked to comorbidities such as ischemic heart disease and pneumonia occurring in the affected patients. Another focus of Gupta and Nayak (2021) was SWI/SNF-related, matrix-associated, actin-dependent regulator of chromatin, subfamily a, member 4 (*SMARCA4*), which is a paralog being the catalytic subunit of the SWI/SNF remodeling

complex, thus helping to regulate chromatin structure.

Regarding viral interactions, large ribosomal subunits from the 40S and 60S proteins were found to interact with viral factors physically. Ribosomal protein L30 (*RPL30*) and ribosomal protein L21 (*RPL21*), identified by Lee et al. [2021], exhibited strong connections to various diseases (Figure 3.9C). *RPL30* was connected to many PRS, including all three COVID-19 genetic risk nodes. In contrast, *RPL21* had additional connections to the comorbidity of ischemic heart disease, often occuring in hospitalized patients with COVID-19.

The genes cullin (*CUL2*) and DNA (cytosine-5)-methyltransferase 1 (*DNMT1*), identified by Gordon et al. [2020], exhibited functions in DNA methylation, maintaining methylation patterns and marking proteins for degradation via ubiquitination, respectively.

Finally, we focus on the similarity graph of genes accumulated from the eight studies on serum or plasma-derived proteome and transcriptome patient data. In four studies, we observed close similarity and, thus, connections to genes and proteins. Among them, eight genes encoded proteasome subunits. These genes exhibited frequent connections to diseases such as MDD, schizophrenia, heart failure, smoking and Crohn's disease. Proteasome 26S subunit, ATPase 1 (*PSMC1*), *PSMC2*, and *PSMA4* were also associated with adipose tissue and the risk of severe COVID-19.

Furthermore, we identified 11 ribosome subunits (seven RPL, three RPS, and one RNF) that exhibited connections to various PRS, predominantly linked to heart failure, MDD, and obesity.

Protein tyrosine phosphatase non-receptor type 6 (*PTPN6*) exhibited associations with one COVID-19 node and the most diseases ($n = 7$), ranging from COPD to hypertension and cerebrovascular disease. The cellular processes it regulates include cell growth and oncogenic transformation. It has been observed to be expressed in B cells during severe COVID-19 [Stephenson et al., 2021]. Cytoskeleton genes were tubulin beta class I (*TUBB*), encoding beta-tubulin, and spectrin alpha, non-erythrocytic 1 (*SPTAN1*), encoding a scaffolding protein. SPTAN1's multi-omic context proved interesting since it was connected to pneumonia and T2D as comorbidities and the tibial artery as a tissue.

Several genes identified across serum and plasma patient data were associated with the cellular response to DNA damage and repair. They included RAD51 recombinase (*RAD51*), bloom syndrome RecQ-like helicase (*BLM*), Erb-B2 receptor tyrosine kinase 2 (*ERBB2*), and mediator of DNA damage checkpoint 1 (*MDC1*). The last gene belongs to the epidermal growth factor receptor family of receptor tyrosine kinases and has been previously associated with the cytokine release storm and SARSCoV-2 infection severity [Khitan et al., 2022]. This gene has gained particular interest in its correlation with obesity and gut microbiome in COVID-19 patients.

We also examined the tissue-specific context of COVID-19–associated genes identified in blood, with a focus on samples from severe patients, regardless of their additional associations with genetic predispositions. Brain-related tissues such as the nucleus accumbens basal ganglia, substantia nigra, and spinal cord were among the most frequently associated tissues, followed by the tibial artery, small intestine, and pituitary (Table 3.2). Potentially triggering cellular processes associated with neurodegeneration and worsening conditions such as Parkinsonism, SARS-CoV-2 has been found in the substantia nigra during the acute infection phase, where it preferentially targets dopaminergic neurons [Bouali-Benazzouz and Benazzouz, 2021]. Lehmann et al. [2021]

described that the small intestine was also impacted in COVID-19 patients. Furthermore, [Frara et al., 2021] documented poorer outcomes in patients with pituitary dysfunction, particularly those exhibiting abnormal endocrine phenotypes such as diabetes and hypopituitarism. COVID-19 patients also exhibited worse outcomes with thrombosis in the tibial arteries [Singh et al., 2021].

Table 3.2: Tissue context of known COVID-19 genes derived from different study types of GWAS, CRISPR, Physical interaction and patient cohorts, detailed in counts and percentage of all tissues. (From Hu et al. [2022])

	Tissue	**Count**	**Proportion of all tissues [%]**
GWAS	Brain spinal cord cervical c-1	27	4.6
	Breast mammary tissue	22	3.75
	Brain frontal cortex BA9	22	3.75
	Cells EBV-transformed lymphocytes	22	3.75
	Heart atrial appendage	22	3.75
CRISPR	Brain nucleus accumbens basal ganglia	4	6.56
	Artery tibial	4	6.56
	Adipose subcutaneous	4	6.56
	Brain anterior cingulate cortex BA24	3	4.92
	Breast mammary tissue	3	4.92
Physical interaction	Brain substantia nigra	9	8.11
	Pituitary	8	7.21
	Small intestine terminal ileum	5	4.5
	Pancreas	5	4.5
	Adipose visceral omentum	5	4.5
Patient	Brain spinal cord cervical c-1	142	4.33
	Breast mammary tissue	127	3.88
	Stomach	126	3.85
	Pituitary	122	3.72
	Cells EBV-transformed lymphocytes	118	3.6

3.3 Path representations predicted novel links in biomedical KGs

Moving from multi-omic integration and embedding towards the explicit modeling of relationships between biological entities, we applied our method BioKGC with adaptation to the specifics of biomedical KGs to the two tasks of function annotation and drug repurposing (Section 2.7). We show how make use of existing knowledge from both fields to establish true but missing links, which can be used for hypothesis generation and provide interpretability.

3.3.1 Functional annotation

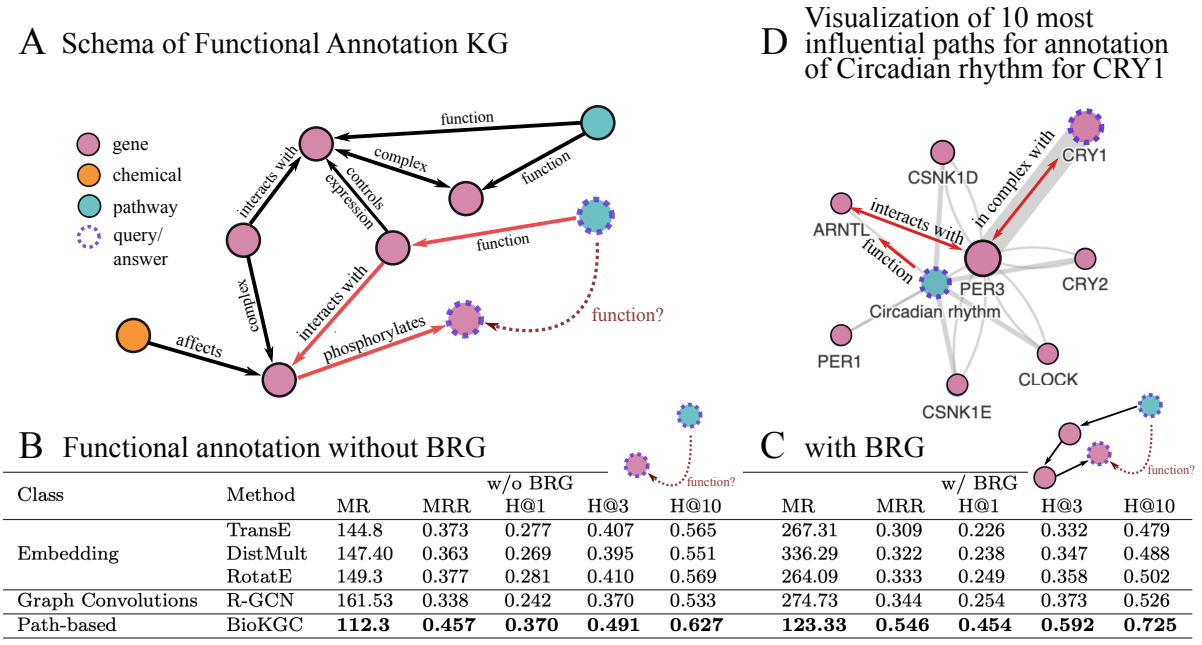
For P-G prediction in the functional annotation task, the BRG we created contained G-G, chemical-gene (C-G), and chemical-chemical (C-C) interactions. It consisted of 30,885 nodes, of which 19,554 were genes and 11,331 were small molecules. The BRG comprised 1,884,146 relations of 13 unique types (Table 3.4), including interactions such as *controls phosphorylation of*, *interacts with* and *chemical affects* (Table 3.3). After data preprocessing that removed KEGG pathways with little annotation and disconnected components in the joint graph, 321 of the initial 333 KEGG pathways were retained. For example, "circadian rhythm" was connected to 8,000 genes (P-G), consisting of 32,367 (of 32,494 initial) assignments. After splitting the supervision edges, we obtained $22,702$ for training, $3,243$ for validation, and $6,487$ for testing (Table 3.3).

In a direct comparison of KGE methods, graph convolutions and BioKGC, our path-based reasoning method, outperformed all other methods, with an average MRR over five model seeds of 0.457 versus an average of 0.371 for KGE and 0.338 for R-GCN in the setting without the BRG. For the Hits@k metric, informative about the proportion of ground truth items captured in the top k predictions, RotatE scored 56.9% in the top 10, compared to 62.7% for BioKGC (Figure 3.11B). When the BRG was leveraged for its biological regulations in message passing, the performance increased only for R-GCN and BioKGC. While the MRR improved from 0.338 to 0.344 for R-GCN, it increased from 0.457 to 0.546 for BioKGC, an 18.3% gain in performance. Within the top 10 predictions, BioKGC could capture 72.5% of the ground truth positives compared to only 52.6% for R-GCN (Figure 3.11C). Interestingly, KGE methods could not leverage the additional information provided by the BRG. Rather than being advantageous for the prediction, performance decreased most severely for the TransE model (from 0.373 to 0.309). When examining each method with five different model seeds, the MRR standard deviation ranged from 0.001 to 0.008, indicating robust predictions for all methods.

Besides its better performance, another advantage of BioKGC over the embedding-based methods is its accessible interpretability, visualizing the most important paths for a given prediction. We used our interpretability tool to explore the most important paths for the predictions. The visualization plot highlights the top 10 most significant paths, ranked by their gradient values. The edge width in the plot represents the frequency with which each edge appears across these paths, providing an intuitive representation of their relative importance. Furthermore, the most important path is highlighted in red, and nodes are colored by their type. For example, we observed the most important path for the functional prediction of cryptochrome circadian regulator 1 (*CRY1*) to the *Circadian rhythm* to be: *CRY1* `in complex with` → period circadian regulator 3 (*PER3*) `interacts with` → basic helix-loop-helix ARNT like 1 (*BMAL1/ARNTL*) before feeding into the `function of` the *circadian rhythm* pathway (Figure 3.10D). The essential

Table 3.3: Summary of the functional annotation dataset: Number of nodes, edges, and relation types in the BRG constructed from Pathway Commons, as well as training, validation, and testing sets obtained from KEGG. (Adapted from Hu et al. [2024])

	Number of nodes	Number of edges	Number of relations
BRG	30,885	1,884,146	13
Training	7,070	22,702	1
Validation	2,348	3,243	1
Testing	3,533	6,487	1
Total count	31,410	1,916,513	14



Class	Method	w/o BRG MR	MRR	H@1	H@3	H@10	w/ BRG MR	MRR	H@1	H@3	H@10
Embedding	TransE	144.8	0.373	0.277	0.407	0.565	267.31	0.309	0.226	0.332	0.479
	DistMult	147.40	0.363	0.269	0.395	0.551	336.29	0.322	0.238	0.347	0.488
	RotatE	149.3	0.377	0.281	0.410	0.569	264.09	0.333	0.249	0.358	0.502
Graph Convolutions	R-GCN	161.53	0.338	0.242	0.370	0.533	274.73	0.344	0.254	0.373	0.526
Path-based	BioKGC	**112.3**	**0.457**	**0.370**	**0.491**	**0.627**	**123.33**	**0.546**	**0.454**	**0.592**	**0.725**

Figure 3.10: **Benchmarking of KG completion algorithms for functional annotation.** A) Illustration of BioKGC leveraging a BRG consisting of genes, chemicals, and cellular pathways and their relations for predicting the functions of genes with their cellular pathway. Functional annotation performance with different KG completion methods, from classical KGE to GCN-based to path-based BioKGC, for link prediction (B) without and (C) with the BRG. D) Visualization of gradients on paths that were important for predicting the link between CRY1 and circadian rhythm. The top 10 paths are shown, with edge width reflecting edge weight, and the path with the highest weight colored red. (Adapted from Hu et al. [2024])

clock circadian regulator (CLOCK) and transcription factors controlling the cellular circadian rhythm ARNTL, upregulate the expression of *PER3* and cryptochrome circadian regulator 2 (*CRY2*) [Gekakis et al., 1998, Olkkonen et al., 2017]. In turn, they form heterodimers to repress the expression of the former, creating a negative feedback loop [Griffin et al., 1999, Kume et al., 1999].

Table 3.4: Detailed breakdown of the relation types in the functional annotation dataset: Number of edges per relation type in the BRG, training, validation, and testing sets. (Adapted from Hu et al. [2024])

	Relation	Count
BRG	Controls-transport-of-chemical	3,741
	Reacts-with	4,063
	Controls-transport-of	7,899
	Used-to-produce	14,747
	Controls-phosphorylation-of	17,660
	Controls-production-of	21,262
	Consumption-controlled-by	22,659
	Controls-expression-of	125,860
	Catalysis-precedes	147,948
	In-complex-with	191,275
	Controls-state-change-of	191,548
	Interacts-with	517,390
	Chemical-affects	618,094
Train	KEGG pathway	22,702
Valid	KEGG pathway	3,227
Test	KEGG pathway	6,438

3.3.2 Drug repurposing

After the proof-of-concept that BioKGC can be used to reason on biomedical KGs with good results, we turned to a more challenging task of zero-shot prediction of drugs for diseases. Here, we utilized the PrimeKG database, which encompasses 129,375 nodes across 10 distinct node classes and 8,100,498 edges representing 27 different types of relationships. The node and edge relations are summarized in Huang et al. [2023].

After removing reverse relations to adapt for BioKGC, ∼5.7 million directed edges were used solely for message passing in the BioKGC model in each prediction setting (Table 3.5). These connections were not drug-disease edges but rather represented protein-protein or disease-disease relationships. In contrast, only edges between drugs and diseases, i.e. indication, contraindications, and off-label use, were used for message passing and supervision. On average, there were ∼33,000 and ∼4,000 edges in the training and validation sets, respectively, in the five examined disease areas: adrenal gland, anemia, cardiovascular, cell proliferation, and mental health-related. The number of testing edges varied greatly between the disease areas. While there were only 303 and 33 were present in the adrenal gland disease area, there were 1,047 and 999 contraindications and indications in the cell proliferation split. These differences stem from the number of diseases within each area (Table 3.5) and can probably also be attributed to the nature of biases in research.

For each disease area, we evaluated the performance of TxGNN and BioKGC in predicting the ground truth drugs given the disease and the relation contraindication and indications. We achieved higher AUPRCs for two of the five disease areas for contraindication and all disease areas for indication (Figure 3.11C). The difference Δ in AUPRC was calculated as BioKGC minus TxGNN; thus, a positive Δ indicates the BioKGC model performed better. In the contraindica-

Table 3.5: The number of diseases per disease area and the number of edges used for the BRG, training, validation, and testing. (Adapted from Hu et al. [2024])

Disease area	Number of diseases	BRG	Training	Validation	Testing (contraindication)	Testing (indication)
Adrenal gland	6	5,728,452	33,063	4,723	303	33
Anemia	19	5,705,775	33,715	4,817	752	88
Cardiovascular	111	5,695,332	30,930	4,419	4,215	453
Cell proliferation	201	5,689,920	33,102	4,729	1,047	999
Mental health	60	5,690,512	33,443	4,778	1,567	355

tion prediction task, TxGNN performed better for The adrenal gland, cardiovascular, and mental health disease areas with Δ of -4.6, -0.2 and -2.0 percentage points. In contrast, BioKGC performed better for anemia and cell proliferation with Δ of 4.8 and 9.0. All Δs were positive in the indication prediction task, ranging from 5.9 to 22.6 percentage points (Figure 3.11D).

3.3.2.1 Cell proliferation split

A detailed breakdown of both models' specificity, F1 score, and Recall@k is illustrated for the cell proliferation disease area (Figure 3.11E). Both models perform well in rating true negative items as negative, with BioKGC exhibiting only slightly higher specificity in the indication setting (0.981 vs. 0.996). The F1 score, which symmetrically represents a model's precision and sensitivity, was 0.330 for TxGNN and 0.415 for BioKGC in the contraindication setting, and 0.183 and 0.393 in the indication setting, respectively. We observed greater performance increases for indication than for contraindication for cell proliferation and all other disease areas. However, the higher variance for BioKGC than for TxGNN is also of interest.

In the cell proliferation split, 178 diseases were identified with known indications, averaging 5.58 indications per disease. For 60 of the 178 diseases, all known treatments were prioritized in the top 10 predictions out of over 7,000 drug candidates. The metric Recall@k describes the top k predictions in a single number.It reflects the proportion of all ground truth items that are included within the top k predictions. For example, at $k = 20$, the Recall@k for indication over across all diseases was 0.619 for BioKGC, indicating that 61.9% of ground truth drugs were included within the top 20 ranked predictions. In contrast, Recall@k was 0.539 for TxGNN. Therefore, BioKGC consistently outperformed TxGNN for low ks (Figure 3.11E) and is equal to TxGNN for $k = 100$ (data not shown).

To provide a single performance metric, the AUPRC was calculated as the average across contraindications and indications for each disease area (Table 3.7). For cell proliferation, the difference was $0.556 - 0.437 = 0.119$, indicating that BioKGC outperformed TxGNN by $0.119/0.437 = 27.3\%$. Similarly, the increase was 17.1%, 14.1%, 25.9%, and 16.9% for adrenal gland, anemia, cardiovascular, and mental health. On average, BioKGC outperformed TxGNN by 20.2% over all disease areas.

ALL in the cell proliferation split

After quantitatively evaluating the performances, we investigated individual predictions of diseases within the cell proliferation split. ALL is a type of cancer that affects the blood and bone

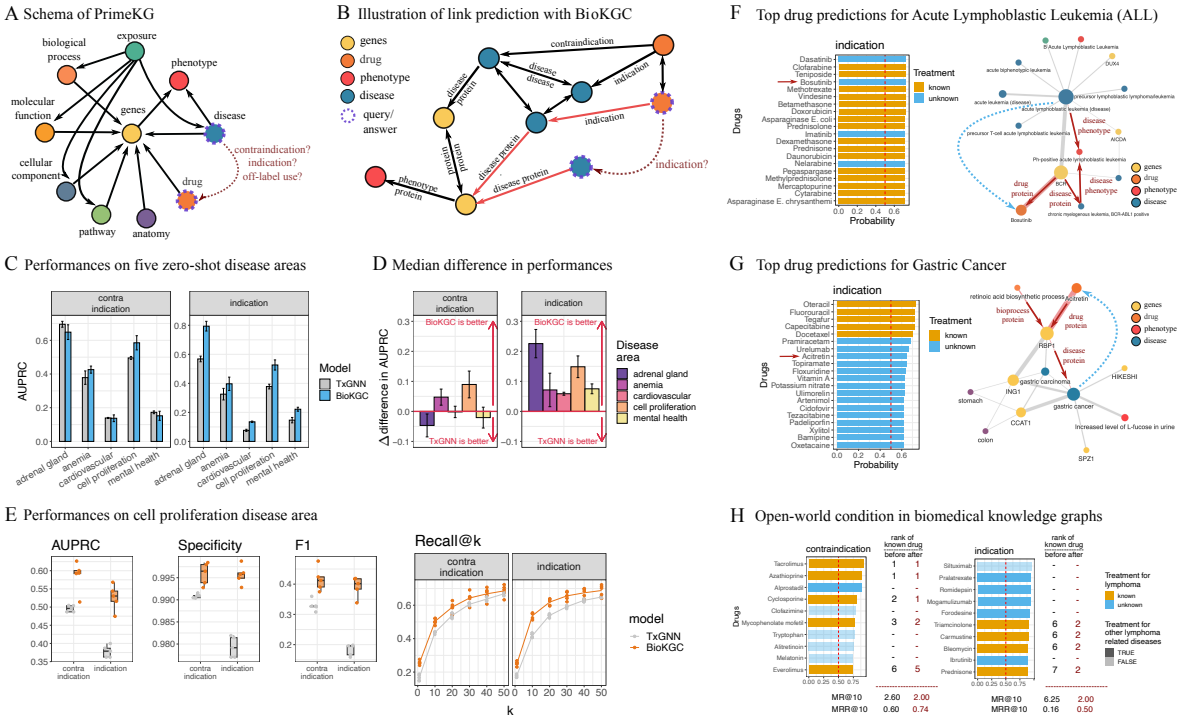


Figure 3.11: **Comparison of BioKGC with the state-of-the-art TxGNN model for predicting drug-disease relations in PrimeKG.** A) Schema of the PrimeKG (left), a multi-modal KG for predicting relations between drug and disease, integrating 10 different biological node types, such as protein, disease, phenotype, anatomy, molecular function, drug, pathway, and exposure, with over five million relations. B) Illustration of BioKGC leveraging paths between drugs and diseases for indication prediction. C) Comparison of AUPRC performances in the zero-shot prediction scenario following the scheme from TxGNN across five disease areas: adrenal gland, anemia, cardiovascular, cell proliferation, and mental health. The AUPRC was computed by comparing all ground truth positive drugs against all ground truth negative drugs for a given disease and then averaging metrics across all diseases within a disease area. As defined by TxGNN, the zero-shot prediction scenario excludes all treatments for disease areas from the training set and removes 95% of edges from a disease to surrounding biological entities, mirroring little molecular characterization. D) Mean differences in AUPRC for each disease area were calculated to compare BioKGC with TxGNN. This metric highlights the relative performance across different disease domains, with a positive Δ reflecting higher performance. E) Performance metrics specifically for the cell proliferation disease area. Recall@k represents the proportion of ground truth edges successfully retrieved within the top k predictions. Top predictions of contraindication and indication for the disease (F) ALL and (G) ovarian mucinous adenocarcinoma within the cell proliferation disease area. Known treatments are included in the ground truth in PrimeKG. Visualization of gradients on paths important for predicting (H) bosutinib for ALL and (I) orafenib for ovarian mucinous adenocarcinoma. The top 10 paths are shown, with edge width reflecting the edge weight, and the path with the greatest weight colored red. (From Hu et al. [2024])

marrow, involving the abnormal differentiation and proliferation of lymphoid cells that enter the bloodstream and interfere with the production of other blood cells and reduce the ability to fight infections [Pui and Evans, 1998]. It is a complex disease with many factors contributing to its development, from genetic syndromes, such as Down syndrome, to environmental factors, such as viruses [Bielorai et al., 2013, Chessells et al., 2001, German, 1997, Sehgal et al., 2010, Shah et al., 2013]. However, chromosomal aberrations are commonly observed, such as the translocation t(9;22) producing BCR-ABL1, a constitutively active tyrosine kinase with the phenotype termed (Philadelphia) Ph-positive ALL [Terwilliger and Abdul-Hay, 2017].

We used BioKGC to predict the drugs associated with ALL. We retrieved the only known contraindication, the drug alprostadil, which ranked first with a probability score of 0.727 (Figure 3.12A), and all 21 known indications within the top 34 predictions (Figure 3.11F). Investigating the top indication predictions, we found the top three known treatments to be clofarabine (0.723), teniposide (0.723), and methotrexate (0.720), and the top unknown treatments to be dasatinib (0.724) and bosutinib (0.721) (Figure 3.11F).

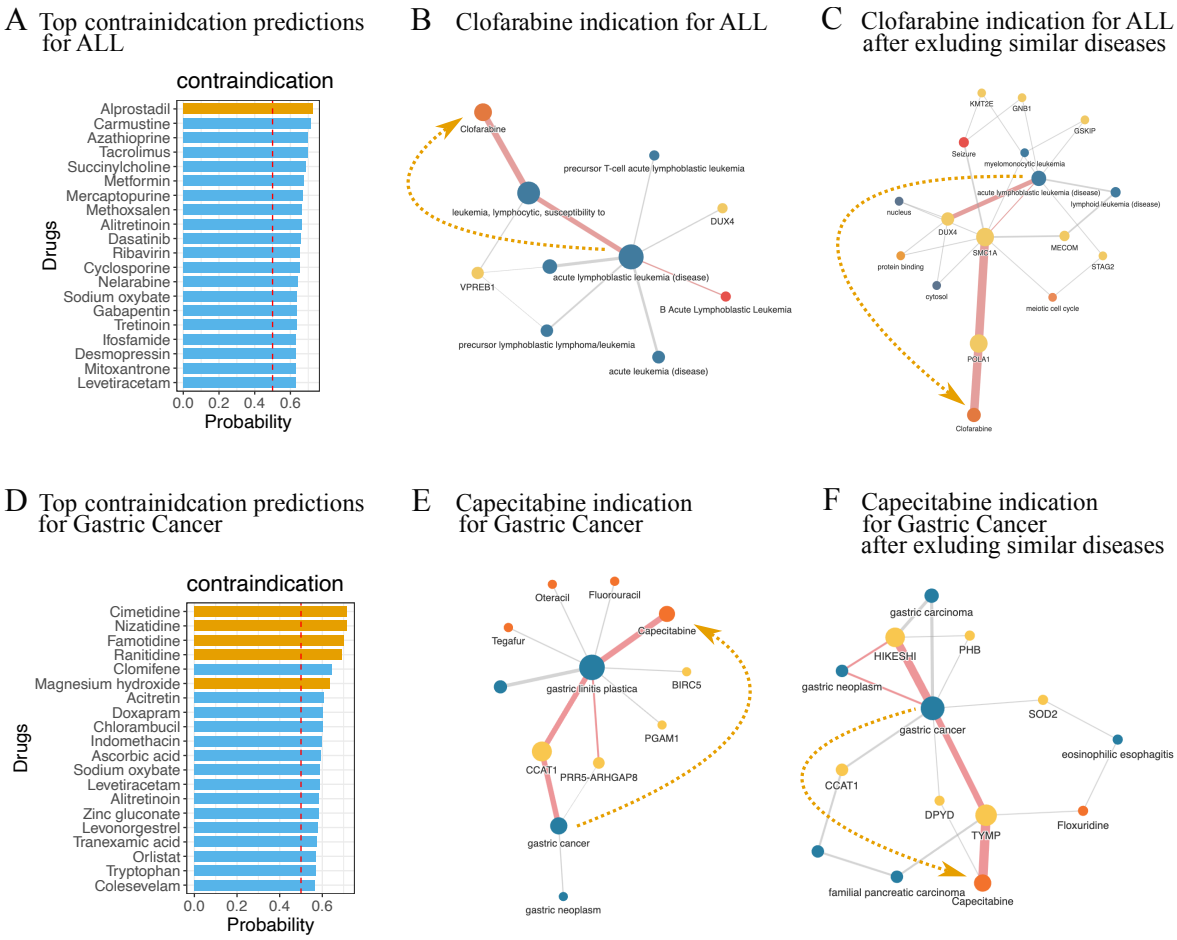


Figure 3.12: **Contraindications and known treatments.** The top contraindication predictions for (A) ALL and (D) gastric cancer, and the visualization of important paths for the prediction of known drug clofarabine for ALL using the (B) training graph and (C) modified graph removing similar diseases. Similarly, visualization of the important paths for predicting capecitabine for gastric cancer using the (E) training and (F) modified graphs. (Adapted from Hu et al. [2024])

Bosutinib is an unknown drug predicted to treat (indication) ALL with a probability score of 0.721 (Figure 3.11F left). With our interpretability tool, which identifies the most important paths for a given prediction, we can focus on the local subgraph. Here, for the prediction of indication bosutinib for ALL, the similarity to other (lymphoblastic) leukemia types was uncovered along with important disease genes activation-induced cytidine deaminase (*AICDA*) and double homeobox 4 (*DUX4*) [Messina et al., 2011, Tanaka et al., 2018, Zhang et al., 2016]. The most important path was from the phenotype *Ph+ ALL* `disease phenotype` → the disease *chronic myelogenous leukemia, BCR-ALB1 positive* `disease phenotype` → the gene BCR activator of RhoGEF and GTPase (*BCR*) → the drug *Bosutinib* via a `drug protein` relation (Figure 3.11F right). Indeed, bosutinib, initially approved in 2012 for the treatment of chronic myeloid

leukemia, is now under investigation for its potential use in ALL [Knox et al., 2024, Mortlock et al., 2017, Varallo-Rodriguez et al., 2015].

Clofarabine is an established treatment documented in the ground truth PrimeKG database. For this prediction, the model identified the disease similarity between ALL and susceptibility to lymphocytic leukemia and leveraged its link to the drug (Figure 3.12B). To investigate the paths the model would select, we excluded links between similar diseases and drugs in the KG, ensuring that disease similarity could not be leveraged for prediction. This time, important nodes that were retrieved were the genes structural maintenance of chromosomes 1A (*SMC1A*), involved in chromosome cohesion during cell division and DNA repair, and DNA polymerase alpha 1, catalytic subunit (*POLA1*), part of the DNA polymerase alpha subunit (Figure 3.12C). Since clofarabine is a purine nucleoside that is metabolized intracellularly to inhibit DNA synthesis [Huguet et al., 2015, Knox et al., 2024], the model detected important components on its mode of action.

Hypothesis generation with BioKGC for treating gastric cancer

To demonstrate the ability of BioKGC to generate hypotheses to test and evaluate in the laboratory, we focused on the example of gastric cancer. Like for ALL, contraindications and indications were retrieved in the top ranks (all five contraindications were in the top 6, and five of the six indications were in the top 5) (Figure 3.11G left). One of the drugs predicted for treating gastric cancer was acitretin, an oral retinoid with a similar structure to vitamin A. It is an indication for treating skin-related diseases, such as psoriasis, inhibiting excessive cell growth and keratinization [Knox et al., 2024]. This drug has not been tested for treating gastric cancer, even though there have been efforts to combine acitretin with clarithromycin for treating cutaneous squamous cell carcinoma due to its ability to induce apoptosis [Lin et al., 2009, Zhao et al., 2021]. Interestingly, in the interpretability plot, all paths from gastric cancer to the drug pass through retinol-binding protein 1 (*RBP1*), annotated with the *retinoic acid biosynthesis process* (Figure 3.11G, right). Recent studies have also shown an involvement of the pathway for treating gastric cancer and provided pre-clinical evidence that the use of all-trans retinoic acid (ATRA) might be beneficial [Guarrera et al., 2023, Nguyen et al., 2016b]. Using the interpretability tool to visualize important paths for predictions between drugs and diseases, researchers can easily verify their plausibility and generate hypotheses that can be further assessed and perhaps validated in the laboratory.

The open-world assumption in biomedical KGs

True to all KGC algorithms, the evaluation metrics of models are based on the closed-world assumption and might not hold in real life. During evaluation, positive triplets in the testing set are generally ranked against all negative triplets. However, when a large portion of the false triplets are actually true (open-world assumption), the ranking-based metrics, such as MR and MRR, report lower performances [Yang et al., 2022]. We believe that this might especially be expected with biomedical KGs because of their high incompleteness, such as due to missing research. For illustration, we focus on the example of the disease lymphoma, which has seven contraindications and 20 indications. However, when we include all lymphoma-related diseases, such as ALL, Hodgkin's lymphoma, and primary central nervous system lymphoma, there are eight contraindications and 70 indications. In the top 10 predictions for lymphoma, the ranks of the known drugs (positive triplets) are lower when ranked against all negative triplets (e.g.,

yielding an MRR of 0.16 for indications). However, when we correct the ground truth positives to include drugs of specific lymphoma diseases, the ranks of the individual triplets rise, and the MRR increases from 0.16 to 0.5 (Figure 3.11H). This illustrates a potential misalignment of evaluation metrics with the power of the model.

3.3.2.2 Custom Alzheimer's Disease (AD) split

AD is a devastating neurodegenerative disorder that is characterized by the accumulation of extracellular amyloid beta and intracellular tau proteins in the brain. These neuropathological changes manifest decades before any clinical symptoms emerge and ultimately lead to synapse loss and brain atrophy, causing clinical symptoms of dementia, such as memory loss and behavioral changes. Despite amyloid and tau being the central disease hallmarks, the exact pathogenic mechanisms driving AD onset and progression remain poorly understood. Emerging evidence suggests the involvement of additional pathways, such as immunoinflammation and bioenergetic dysregulation [Batra et al., 2023, Bellenguez et al., 2022, Heppner et al., 2015], which may offer promising therapeutic targets. Presently, US Food and Drug Administration (FDA) has approved only two disease-modifying treatments and five symptomatic therapies for Alzheimer's disease (AD), none of which provide a cure. To evaluate the potential of BioKGC in addressing complex and heterogeneous diseases, we trained the model on a dataset split specifically designed for zero-shot prediction, focusing on a custom-defined AD area split. Here, we used TxGNN code for disease evaluation and excluded all treatments for AD and closely related diseases (Pick disease, AD without neurofibrillary tangles, and Lewy body dementia and dementia) (Table 3.6). We then investigated the top 20 drug predictions for AD.

Seven of the eight drugs included in PrimeKG and four of the seven FDA-approved treatments for AD were successfully retrieved among the top 14 predictions (Figure 3.13A). These include several cholinesterase inhibitors (ipidacrine, donepezil, tacrine, rivastigmine, and galantamine) commonly used to reduce neuropsychiatric symptoms [Deardorff et al., 2015]. Epicriptine, a nootropic drug with an undefined mechanism of action, and acetylcarnitine, which plays a functional role in the β-oxidation of fatty acids [Jones et al., 2010], were also identified. Among the known drugs assigned low probability scores was pramiracetam, which ranked 344. This drug is commonly used to address cognitive impairment linked to aging and dementia [Malykh and Sadaie, 2010]; Additionally, FDA-approved treatments such as memantine (ranked 412), an *N*-methyl-D-aspartate receptor antagonist [Deardorff et al., 2015], and the recently approved monoclonal antibodies lecanemab (ranked 2250) and aducanumab (ranked 2216) were not assigned high probabilities [Budd Haeberlein et al., 2022, Van Dyck et al., 2023].

Of note, two drugs currently undergoing clinical trials appeared among the top 20 predicted indications. One of them, nicotine, a nicotinic acetylcholine receptor agonist, is being evaluated in a Phase II clinical trial (NCT02720445) for its potential to improve cognition. The second was bupropion, an *N*-methyl-D-aspartate receptor antagonist, currently being tested as part of the drug AXS-05 in two Phase III clinical trials (NCT05557409 and NCT04947553) [Cummings et al., 2023] for its potential to alleviate agitation associated with with AD. An analysis of the interpretability graphs reveals that both predictions are linked to the brain-derived neurotrophic factor (*BDNF*), a gene essential for synaptic maintenance and plasticity in the brain [Kowiański et al., 2018] (Figure 3.13C). Synaptic plasticity is pivotal in AD [Styr and Slutsky, 2018], with research indicating lower BDNF levels in both the blood [Ng et al., 2019] and brain [Wan et al.,

2020] in patients with AD and linking higher brain BDNF levels with slower cognitive decline [Buchman et al., 2016] in older adults. In addition, BDNF's involvement in downregulated pathways is underscored by analyses of brain gene co-expression networks ($P_{adj} = 6.93e-17$, adatlas.org [Wörheide et al., 2021]). Both predicted drugs, nicotine and bupropion, have demonstrated an ability to elevate serum BDNF levels [Jamal et al., 2015, Tafseer et al., 2021], providing a functional hypothesis for their mechanism in the context of AD.

Another promising candidate is everolimus which has been predicted with high probability, a rapamycin analog and selective inhibitor of the mechanistic target of rapamycin kinase (MTOR) signaling pathway. This pathway has been implicated in both normal and pathological aging processes, positioning it as a promising target for therapeutic intervention, especially during the early stages of the AD [Mannick and Lamming, 2023]. Rapamycin is currently being evaluated as potential disease-modifying therapy in clinical trials, including a Phase II trial (NCT04629495) and a Phase I trial (NCT04200911), both involving older adults with mild cognitive impairment or early-stage AD [Cummings et al., 2023]. Furthermore, rapamycin exhibits beneficial effects on tau and amyloid burden in AD mouse models [Kaeberlein and Galvan, 2019]. While structurally similar to rapamycin, everolimus exhibits favorable clinical pharmacokinetics, including improved bioavailability and tissue distribution [MacKeigan and Krueger, 2015]. As a result, everolimus could be a promising candidate for targeting the hyperactivated MTOR pathway in AD.

Table 3.6: Zero-shot prediction scenario for AD: the diseases that were considered for this analysis and number of indication and contraindication drugs in custom data split generated following the "disease evaluation" code from TxGNN code. (Adapted from Hu et al. [2024])

Disease	ID	Number of contraindications	Number of indications
AD	28780	56	8
AD without neurofibrillary tangles	83960	2	7
Lewy body dementia	29296	2	0
Pick disease	28473	27	7
Dementia	37573	28	3

Table 3.7: Performance in five different disease areas. The median metrics are summarized as the average across contraindications and indications. (Adapted from Hu et al. [2024])

Adrenal gland

	Recall @20	AP @20	MRR @20	F1	AUPRC 1:1	AUPRC
TxGNN	**0.591**	0.728	0.720	**0.616**	0.585	0.632
BioKGC	0.564	**0.901**	**0.917**	0.393	**0.588**	**0.729**

Anemia

	Recall @20	AP @20	MRR @20	F1	AUPRC 1:1	AUPRC
TxGNN	0.445	0.458	0.513	0.236	0.666	0.339
BioKGC	**0.503**	**0.496**	**0.526**	**0.311**	**0.678**	**0.412**

Cardiovascular

	Recall @20	AP @20	MRR @20	F1	AUPRC 1:1	AUPRC
TxGNN	0.102	0.173	0.198	0.059	0.604	0.108
BioKGC	**0.17**	**0.230**	**0.253**	**0.075**	**0.613**	**0.139**

Cell proliferation

	Recall @20	AP @20	MRR @20	F1	AUPRC 1:1	AUPRC
TxGNN	0.535	0.493	0.557	0.260	0.823	0.442
BioKGC	**0.605**	**0.625**	**0.662**	**0.387**	**0.829**	**0.561**

Mental health

	Recall @20	AP @20	MRR @20	F1	AUPRC 1:1	AUPRC
TxGNN	0.166	0.230	0.267	0.112	0.617	0.156
BioKGC	**0.211**	**0.261**	**0.285**	**0.123**	**0.576**	**0.188**

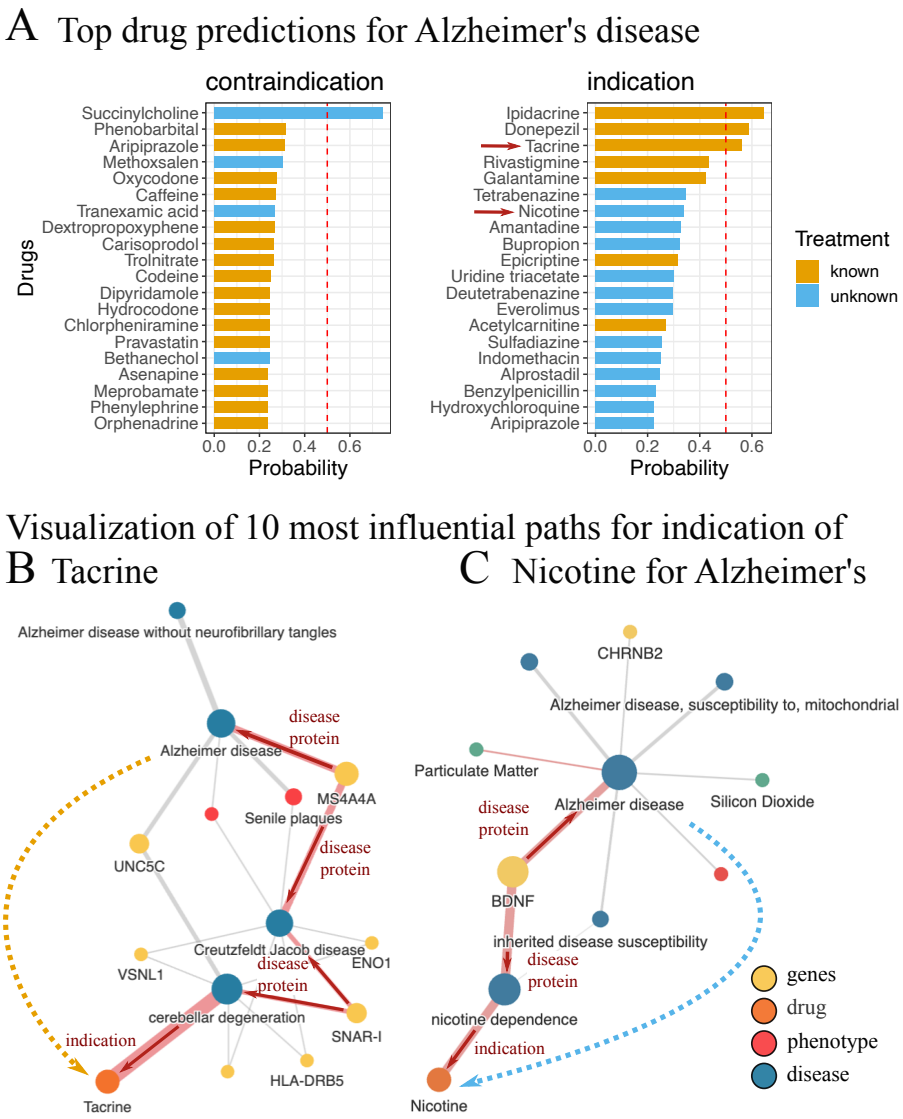


Figure 3.13: **BioKGC predictions in a custom data split for AD.** The top predictions of contraindications and indications for custom AD. The data split was obtained using TxGNN's disease evaluation code. The ground truths were known treatments in PrimeKG. Visualization of the gradients on paths important for predicting (B) known tacrine and (C) unknown nicotine for AD. The top 10 paths are shown, with edge width reflecting edge weight, and the path with the greatest weight colored red. (Adapted from Hu et al. [2024])

Chapter 4

Discussion and Outlook

Biological systems can naturally be modeled in networks with biological entities as nodes and their interplay as edges. Especially the ability of network models to comprehensively capture complex interactions is of importance to gain deeper understanding of the system, e.g. that give rise to properties of health and disease. One of the main challenges in network biology revolves around the integration of diverse data types, such as sequencing techniques and personal clinical data, originated from patient and population cohorts, into so-called multi-omic or multi-model networks that are specific to the disease system under study. Further, there is the need for efficient tools to explore these heterogeneous networks to elucidate the context of biological entities, such as genes and diseases, and their interplay. In addition, research aims to predict links between missing entities yet true, using accumulated knowledge from previous research to close gaps in knowledge, coupled with good interpretability options to verify for biological plausibility.

In this thesis, I made contributions tackling all of these challenges in network biology. I co-developed a multi-omic integration framework, KiMONo, to obtain a condition-specific network derived from patient cohort data that retains statistical associations over regularized regression (Section 2.5). Further, I proposed a node embedding based framework that utilizes random walks and shallow embedding from the natural language processing field for efficient exploration of non-linear associations in these complex heterogeneous networks (Section 2.6). While these two approaches model the interplay between biological entities as unimodal relations, there is great potential in explicitly modeling the diverse relations. Distinction, e.g. between "activation" and "inhibition" of a gene can be made, which is necessary to understand the underlying molecular mechanism of action when performing knowledge completion. Thus, finally, I introduced a graph neural network based tool to predict missing links in biomedical KGs, named BioKGC (Section 2.7). In the following, each section is summarized and ideas for future methodological improvements and extension are provided. This chapter is based on and partly identical to [Hu et al., 2022, 2024, Ogris et al., 2021].

4.1 Summary of multi-omic integration and analysis of MDD

For the first part of this thesis, we proposed a method (KiMONo) that aids the joint analysis of multiple levels of sequencing data by constructing a condition-specific network (introduced in

Section 2.5). The framework integrates prior biological knowledge, reducing complexity through feature preselection based on experimentally validated PPIs, and then uses statistical models to identify dependencies using sparse-group LASSO penalty in a linear multivariate regularized regression framework. After integrating different data modalities, the resulting condition-specific multi-omic network is analyzed using classical network statistics to identify important nodes.

To showcase the power of our KiMONo method, we applied it to the complex disease of MDD, which is influenced by many internal and external factors (results in Section 3.1). We demonstrated that KiMONo could identify a substantial portion of eQTL and eQTM genes (34.5% and 26.7%, respectively) that were previously detected by MatrixEQTL using pairwise tests. Moreover, when constructing regulatory networks with all features from various omic levels, we uncovered additional associations that complemented the findings of MatrixEQTL. These associations only became apparent when considering the broader context of omic interconnections and crosstalk. Notably, across all the top hits in the MDD dataset , we observed the relationships to genes and methylation sites linked through second-order connections. For example, we identified the gene SLC39A11 as both an eQTL and an eQTM gene connected to the SNP rs1493550 and the methylation site cg26124719. Our findings emphasize KiMONo as a powerful approach for uncovering these long-distance and indirect relationships while establishing comprehensive regulatory networks. In addition to considering second-order links, we demonstrated the advantages of using multivariate models derived from diverse omic layers. These models revealed intricate interactions between gene expression, genetics, and methylation patterns that remained undiscovered by pairwise models. We could clarify these connections by carefully addressing and accounting for the residual effects of all features except the one of primary interest.

In summary, we demonstrated that KiMONo is a versatile approach for generating fully integrated and holistic multi-level networks, effectively capturing the data-supported interactions across omics levels. Applying KiMONo to MDD revealed that central nodes within the inferred multi-omics disease networks were essential to its pathophysiology. Therefore, the comprehensive networks derived through KiMONo can be valuable tools for uncovering key regulators in complex diseases.

Besides the systematic analysis of robustness regarding the noise levels in the data and sample size, the framework was examined in detail regarding the missingness in the number of features by Henao et al. [2022]. Here, the authors regarded the variance explained with varying degrees of missingness in the features. Among the investigated models were a few that could deal with missing data in random or block missingness, such as knnSGLasso, S(A)Lasso, G(A)Lasso, HMLasso, CoCoLAsso, and BDCoCoLasso. More ideas for future work are discussed jointly with the framework for network embedding (Section 2.6) in Section 4.2.2.

4.2 Summary and outlook for network embedding in biology

In Section 2.6, I introduced a framework combining network inference and embedding to efficiently integrate and analyze multi-omic information. The first step of the framework relies on the method KiMONo described previously (Section 2.5) to generate a multi-modal network. In the second step, the network was embedding using random walks and a shallow learning approach to obtain low-dimensional representation of each node, which can be used to efficiently query the

relationship between nodes as defined by vector similarity.

We used this framework to analyze COVID-19 genes using pre-pandemic population cohort data from the GETx project, including expression data from diverse tissues and genomic and phenotypic information for over 900 patients across 50 tissues (results in Section 3.2). First, I created a multi-modal network where genes were associated with other data modalities, such as tissues, phenotypes, COVID-19 comorbidities, and PRS related to COVID-19 severity and hospitalization, through multiple regression. Second, the network was embedded into a latent space and each node was represented by a low-dimensional vector using DeepWalk. Third, we used this representation to efficiently analyze node relationships, linking COVID-19 genes to diseases and tissues, assuming that associated nodes would be nearby within the embedding space and exhibit high similarity scores. We found strong connections between COVID-19 genes and diseases such as ischemic heart and cerebrovascular diseases. Notably, PTPN6 was associated with multiple comorbidities and genetic predisposition to COVID-19, indicating its importance in severe cases. Our approach contextualized COVID-19 genes within tissue and disease frameworks, highlighting the significance of known and novel genes in disease severity.

In step one, the multi-modal network derived from the GTEx pre-COVID-19 population cohort exhibited high quality, yielding high gene model performances, as evidenced by the R^2 values. Additionally, we ensured the embeddings' quality by optimizing the parameters to maximize the variance of the resulting node pair cosine similarity distribution. The low standard deviation of similarity scores between any two vectors across the 100 runs reflected the robustness of the obtained embeddings, further validating the reliability of the embedding. We thoroughly validated the network embedding, based on tissue and disease similarity. Tissue-specific genes were significantly enriched in the set of genes with the k highest similarity scores to the respective tissue compared to those with the k lowest (with k = [50 100, 200, 300, 400, 500]). Our method successfully identified the most important factors for each tissue. Furthermore, the embedding was validated by its ability to reproduce disease-related factors, including those associated with ischemic heart disease, T2D, MDD, chronic kidney disease, and cancer. Interestingly, the top similar nodes originated from the n-hop neighborhood in the initial KiMONo network.

This means our approach accomplished two essential tasks: (1) it prioritized significant factors from the multitude of connections in the original network, and (2) it identified factors connected over one, two, and three hops. With the embedding, we could examine the relationship between any two nodes in the entire network context, not restricted to only directly connected nodes. The information gained beyond the original network was attributed to the embedding's training, which considered sequences within random walks, effectively incorporating the larger context of each node into the embedding process. This comprehensive approach enhances our understanding of complex biological relationships and contributes valuable insights to disease research.

After proving the validity of the embedding space, we comprehensively explored the multi-modal context of previously identified COVID-19 genes, uncovering patterns from different experimental types. Each experimental approach captured distinct information about the genes' functions and associations. For example, networks derived from genes identified through physical interaction experiments shed light on the roles of ribosomal proteins. They highlighted ribosomal proteins' roles, while CRISPR studies emphasized proteases and genes linked to COVID-19 comorbidities, such as *SMARCA5* and *TRAF3*. Patient blood data showed associations with proteases, ribosomes, DNA repair, ubiquitination, and cytoskeletal processes. *PTPN6* stood out in the

patient blood data, exhibiting associations with multiple comorbidities and playing a central role in COVID-19 when other comorbidities, such as hypertension or COPD, were present. Our analysis also revealed the involvement of various tissues, including the brain and small intestine, in COVID-19.

The framework has certain limitations concerning the prior used to establish PPI links. We analyzed around 7,000 genes mapped to the prior, but it could be extended to cover all human protein-coding genes. Additionally, our PRS was limited to GWAS studies, focusing on diseases most relevant to COVID-19. However, other diseases could be included in a cross-tissue cross-disease cohort to expand the analysis.

The strength of our approach lies in the large cohort size, enabling contextualization of the disease's impact across different tissues at a population level. The substantial number of samples provided statistical power, particularly for brain tissues. Notably, our study represents the first analysis of the genetic predisposition to COVID-19 using a pre-COVID-19 population cohort encompassing multiple tissues while considering genetic predisposition, developed diseases, and phenotypes. This comprehensive approach allowed us to understand the complex disease on multiple layers, from genetics to comorbidities influenced by environmental factors, leveraging complementary information.

4.2.1 Extension of embedding space validation

One challenge we faced was validating the embedding space for biological meaningfulness. In our study, we successfully validated it over tissue and disease-specific genes. However, it might also be achieved, for instance, over the strength of module detection of protein complexes [Forster et al., 2021]. We advise researchers to define a way for embedding validation right in the beginning, alongside study design.

4.2.2 Extension of network assembly methods

Alternative statistical models can be considered in addressing multicollinearity and non-linearity in the data to improve network assembly. When features are correlated, least squares estimates may be biased [Zou and Hastie, 2005], leading to poor performance in some biological contexts [Sun et al., 2020]. Techniques such as principal component regression (PCR), which regresses on uncorrelated PCs, can help handle multicollinearity [Jolliffe, 1986]. PCR can also be seen as a dimensionality reduction method if only a subset of PCs with the highest variance are used. Additionally, while sparse-group LASSO and PCR capture linear dependencies, non-linear models can also be explored due to the complexity of biological systems. Random forest models can be a simple alternative, allowing for network assembly by setting thresholds on feature importance derived from permutation or impurity reduction [Breiman, 2001].

4.2.3 Extension of embedding methods

While we used the DeepWalk algorithm [Perozzi et al., 2014] for network embedding, and it captured meaningful relationships between nodes, there are extensions to the simple algorithm that could be explored further. For example, in the case of node2vec [Grover and Leskovec, 2016], the random walks are biased towards breadth-first-search or depth-first-search, making it possible to either focus more on the local neighborhood or sample from the global structure. Chen et al. [2019] used this algorithm to identify oral cancer-related genes. Other node embedding techniques exist besides node2vec (Section 2.4.3.2). Qiu et al. [2018] summarized and unified them into the matrix factorization framework.

Two notable embedding methods are metapath2vec and metapath2vec++, developed to specifically tailor heterogeneous networks [Dong et al., 2017]. Instead of treating all node and edge types identically in random walks, meta-paths predefine the traversal sequences through the network regarding node and edge types. A meta path could be "genetic" - "gene expression" - "tissue" - "disease state", enforcing the order of node types to visit. These metapath-based random walks generate node embeddings in a heterogeneous skip-gram model to construct the heterogeneous neighborhood. These methods could be especially meaningful in biomedical graphs that suffer from modalities with different information content and a highly biased number of nodes, albeit with heavy usage of domain knowledge in defining meta paths. Besides network embedding methods using shallow neural networks, there is also a range of graph DL methods, such as GraphSAGE [Hamilton et al., 2017a].

4.3 Summary and outlook for KG reasoning

As detailed in Section 2.7, I introduced a method to perform link prediction in biomedical KGs, closing gaps in our understanding of biological systems. While the previous frameworks operated on graphs that are based on one type of relation (statistical association obtained regularized regression models), we now turn towards modeling the type of relations between entities in biological knowledge graphs explicitly. This allows for a more detailed investigation and prediction of relations, such as to distinguish between "activation" and "inhibition" or "indication" and "contraindication". Our model, BioKGC, was adapted from NBFNet, a message-passing GNN which learns path representations between two nodes for the prediction of missing links, taking specific relation types into account. BioKGC was adapted to suit the requirements of biomedical KGs in sampling appropriate negative samples and the usage of background information informing predictions. I showcased the method on the proof-of-concept applications of predicting links between genes and their functions for functional annotation and identifying therapeutic opportunities by linking drugs and diseases in a challenging zero-shot prediction scenario (results in Section 3.3). Notably, I also illustrated the strength of BioKGC's interpretability tool to visualize important paths for a prediction, allowing for better understanding, uncovering training or data biases, and evaluating the predictions for biological plausibility.

4.3.1 Functional annotation

For the functional annotation task, which constitutes a proof-of-concept application of reconstructing the KEGG pathway database, we benchmarked BioKGC against various methods, including the classical KGE methods of TransE, DistMult, and RotatE. Furthermore, as a representative of methods using a GCN, we compared BioKGC against R-GCN. In summary, BioKGC outperformed KGE and R-GCN based on learning node embeddings for prediction. In the setting without the BRG, the relative gain in performance was 23.2% over the average KGE model and 35.2% over R-GCN. Leveraging the BRG provided a relative gain of 69.9% and 58.7%, respectively. Since all other methods could not effectively leverage the additional information supplied over the underlying biological regulation for predicting the pathways and genes, BioKGC achieved the best performance.

We hypothesize that BioKGC outperformed embedding methods due to its ability to prioritize relational paths between entity groups of interest rather than treating each triplet (G-G, C-G, C-C, P-G in our case) with equal importance, as embedding methods typically do. This distinction becomes crucial when dealing with a KG containing appreciable noise. Here, the KG may include additional triplets that may not prove helpful, such as those found in the BRG, which contains numerous links not essential to our specific task. Remarkably, leveraging the underlying BRG in BioKGC improved the prediction of gene functional annotations. Moreover, BioKGC enabled us to visualize predictions by examining the gradients along the paths and identifying the prediction's most influential nodes and relations. This interpretability feature allows results to be verified and explained, which embedding methods lack.

4.3.2 Drug repurposing

For a more exciting and challenging application of predicting drugs for treating diseases, we applied BioKGC to a zero-shot prediction scenario defined by TxGNN, a state-of-the-art method tailored to this application. Training the TxGNN model is divided into pre-training and fine-tuning steps. During pre-training, all edges are used to learn node embeddings, giving equal importance to all. However, only triplets with relations of interest (i.e. indication, contraindications, and off-label use) are used during fine-tuning. This approach ensures extra focus on the nodes of interest, especially drugs and diseases. In addition, disease nodes with little molecular characterization are enhanced with a disease similarity auxiliary embedding in a gated manner, depending on the degree of the nodes. In BioKGC, we made non-drug-disease triplets available to the model as a BRG for message passing but not for supervision, allowing us to focus directly on relations of interest. Furthermore, BioKGC did not require learning auxiliary embeddings to enhance the nodes. Provided some paths connect the disease entities and target drugs, meaningful predictions could be made over the relations.

BioKGC outperformed TxGNN in all five zero-shot disease areas: adrenal gland, anemia, cardiovascular, cell proliferation, and mental health-related disease. The average performance increase (AUPRC) across all areas was 20.2%, demonstrating the power of path-based reasoning methods, especially in predicting indications. In addition, BioKGC achieved higher Recall@k values than TxGNN. Overall, 61.9% of the known treatments were recovered by BioKGC at $k = 20$, while only 53.9% were recovered by TxGNN. In conclusion, our model was notably better at

prioritizing positives among k-top predictions (especially at low k, such as $k < 20$).

Since prioritization is especially important in biology, using BioKGC, fewer predictions must be verified for their biological plausibility in practice for hypothesis generation or experimental validation. However, the variance in performance was slightly higher for BioKGC than for TxGNN, likely due to more stable predictions when auxiliary node embeddings were supplied over a disease similarity learning module in TxGNN. In addition, BioKGC relies on the exact edges available to learn path representations after removing 95% of the edges of disease nodes to other biological entities during the data split.

The interpretability tool based on gradients along the path provides insights into details of drug-disease predictions, making validation (e.g., bosutinib for ALL) and the generation of hypotheses possible (e.g., acitretin for gastric cancer or nicotine and bupropion for AD). We showed that the BioKGC model retrieved the most important paths for the prediction, capturing meaningful associations. Also notable is the supply of a new custom graph during inference. While the BioKGC model learned similarities between diseases and relied on them for the prediction, we provided a subgraph where similar diseases to ALL were removed during inference. Here, we discovered that the most important path for the prediction passed through essential proteins in the mode of action of the clofarabine, namely SMC1A and POLA1, related to DNA synthesis. Using a different graph during training and inference is another advantage of path-based compared to embedding-based methods. This difference would be especially impactful in an inductive setting, where a new graph with different nodes is supplied. Our experiment providing a subgraph shows the flexibility to elucidate the interpretability when excluding similar diseases and hints at the potential of a fully inductive KG reasoning. The requirement is that all nodes for inference are connected to the graph and no new relations appear in the prediction queries. We look forward to benchmarking the fully inductive reasoning task in future studies.

4.3.3 Bias from training data and identification of most informative triplets

The limitations of BioKGC include the potential encoding of biases in the training data. The AD example illustrated that almost all known treatments were retrieved in the top predictions. However, many unknown but FDA-approved drugs in the PrimeKG database were not retrieved. Upon investigation, none of these drugs (e.g., pramiracetam, memantine, and lecanemab) had disease indications in the database. Thus, the model could not learn from their connections. Consequently, none were recovered as potential treatments. While the predictions for the known symptom-treating drugs were made through neuropsychiatric-related diseases, emphasis on the molecular interaction may uncover more disease-modifying treatments. Future improvements could exclude message passing over dominant relations such as indication but focus on molecular interactions.

In our examples, BioKGC mostly retrieved nodes of the type "exposure," "phenotype," "gene," "disease," and "drug" with their relations in the local subgraph, such as "disease-disease," "disease-indication," and "disease-protein." They were most informative to the predictions. Future work could focus on ablation studies to systematically evaluate informativeness, removing triplets of one relation type at a time. Moreover, while the learned paths are meaningful and sufficient for treatment prediction, little is revealed about the specifics of the molecular mode of action, such as the subsequent signaling cascade after drug intake. Improvements can be made

to the data used for training for further investigation of biomedical KG for drug prediction. We propose 1) the sparsification and removal of certain dominant edges and uninteresting edges, and 2) the addition of fine-grained information on the molecular interactions on the protein level, such as replacing the PPIs with specific relations (e.g., from Biogrid detailing "phosphorylation," "ubiquitination," "upregulation," and "downregulation").

4.3.4 Misjudgement of model capacity under open-world assumption

A general note is that under the open-world assumption, the metrics reported to evaluate the models' performance might not be entirely consistent with their strength, which usually must be measured on the complete KG. All known positive ground truth items are ranked against unknown triplets in ranking-based performance metrics. However, some positive items missing in the KG could rank higher than known test answers, making the rankings of the latter drop. When recognizing correct answers, the performance might not increase proportionally with the strength of model [Yang et al., 2022]. The authors investigated the degradation and inconsistency of the performance metric MRR for cases of high incompleteness of the links in the KG. They discovered that issues occur primarily when a correlation exists between missing links and certain entities. These issues are relevant for biomedical KGs, where 1) there are many gaps in our knowledge about biological processes, and 2) there is bias in the accumulated knowledge regarding specific fields. Missing facts are not uniformly distributed but often related to certain entities or entity types. For example, not all diseases have been studied uniformly, with some receiving more attention than others, sometimes merely due to their prevalence or complexity.

4.3.5 Improvement of the biological regulatory graph

Future research could focus on tailoring the KG used for drug-disease prediction to include more informative sources. Our interpretability tool revealed the model's biases in leveraging certain relation types for prediction. The most important paths traversed for treatment prediction frequently led over similar diseases and their drug annotations. While these relations are informative, they have limited explanatory power over the molecular drug effect. We counteracted this issue by removing similar diseases from the inference graph, uncovering the mode of action of drugs. However, this issue of preferring drug-disease relations for prediction should be addressed early on. We recommend training the model on edges with fewer drug-disease edges, which can be achieved by augmenting the data or the model. In addition, fine-grained relations (e.g., from the BioGRID Biochemical Activity panel, such as "phosphorylation", "methylation", "sumoylation", "upregulation", and "downregulation") could replace generic relations such as "protein-protein interaction" to fully detail the biological processes [Stark et al., 2006].

Another future perspective of KG reasoning is the use of condition-specific graphs. The KGs we used detailed all possible interactions derived from a wealth of experiments and literature. However, they are not specific. Instead, we suggest a tissue-specific co-expression graph, for instance, to predict functions specific to each tissue. Alternatively, disease graphs built on patient sequencing information could be used for personalized drug predictions. One effort to create customized biomedical KGs has been initiated by BioCypher, harmonizing molecular alongside ontology, genetics, tissue, patient, disease, and drug data [Lobentanzer et al., 2023].

4.3.6 Future extensions from node features, hypergraphs, and graph foundation models

Instead of sub-setting a KG to create a tissue-specific network, as proposed in the paragraph above, another approach would be to use all nodes of the graph and introduce the tissue-specific expression over the node features. This would allow the addition of experimental sequencing data to further inform predictions. Another interesting future extension is the reasoning on hypergraphs with edges connecting more than two nodes, such as the case with polypharmaceutical effects when two drugs are used together, causing an unwanted effect. Lastly, reasoning on biomedical KGs could generally be used to inform large language models to counteract the generation of coherent yet factually incorrect outputs, especially in the biological domain [Pan et al., 2023, Zhang et al., 2024].

Another interesting development is the rise of foundation models, which are trained on broad data in a self-supervised manner and adapted to a range of specific tasks using transfer learning [Bommasani et al., 2022]. They have applications in fields such as language, vision, and robotics. Galkin et al. [2024] recently made the first effort in a graph foundation model for KG reasoning. While inductive inference only generalizes to graphs with unseen nodes while maintaining the same relation types, the authors showed how to transfer across graphs containing arbitrary entity and relation vocabularies. Eventually, creating a foundation model for the biomedical domain trained on the accumulated knowledge across fields would be useful for many specific downstream tasks. A first effort towards foundation model in biomedical domain is given by BioBridge [Wang et al., 2024], where the authors bridge unimodal learned models to unify into one model. This model learns a transformation from one data space to another, facilitating multi-modal behaviour.

4.4 Conclusions

The contributions of this thesis showcase graph ML as a powerful tool to model and understand biological data. Network biology is essential in integrating multiple omics data levels to uncover intricate biological interactions and identify key players in complex diseases such as MDD. In the first contribution of this thesis, we showed how to perform multi-omic integration retaining statistical associations from multi-variate regularized regression models suitable for the $p \gg n$ problem, where the number of variables (p) greatly exceeds the number of samples (n). This contribution addressed the (i) and (ii) challenges of network biology (Section 1.1) by creating a condition-specific network using KiMONo which was derived from a prior containing all possible interactions based on heterogeneous patient data. Based on patient/population cohort quantitative sequencing data, only statistically relevant relations between nodes are retained. Then, key players in health and disease were uncovered over classical network analysis tools to identify the most important genes for the complex disease, MDD.

The second contribution of this thesis addressed the challenges (i), (ii) and (iii) to analyze integrated multi-modal data of condition-specific networks efficiently. This was achieved using network embedding techniques to capture meaningful biological relationships, thereby elucidating the multi-omic context of the disease nodes, such as the heterogeneous disease, COVID-19. We were able to prioritize significant factors from the multi-modal context of a node that were not restricted to only the 1-hop neighborhood. We attribute this information gain beyond the original

network to learning the embeddings based on random walks, which can incorporate a larger context for each node in the embedding process. In addition, the direct similarity operations directly based on nodes' embeddings provide an efficient and straightforward contextualization of a biological entity, such as COVID-19-related genes, allowing the understanding of its multimodal role.

While the first two contributions modeled the statistical interactions between entities, the usage of knowledge graphs and approaches to explicitly take diverse relations into account facilitates the fine-grained understanding of interplay going beyond unimodal associations, thereby addressing the challenge (v). Reasoning on biomedical KGs that exhibit detailed relations, such as "in complex with", "phosphorylation", "indication", accelerates research by learning from existing knowledge to predict new links, fueling hypothesis generation on functional annotation and therapeutics prediction, thereby addressing the challenge (iv) to complement incomplete knowledge. Further, a mechanism of action can be proposed to uncover biases and verify biological plausibility, addressing the challenge (vi) to improve interpretability. However, as discussed in Section 4.3, future work should include condition-specific networks, such as those derived from tissue or disease specificity (challenge (i)).

In summary, graph ML is a powerful tool to model biomedical systems - to integrate heterogeneous data, embed entities meaningfully, and reason on KGs for new hypothesis generation - that will continue to develop and substantially improve our understanding of complex biological systems.

Supplementary

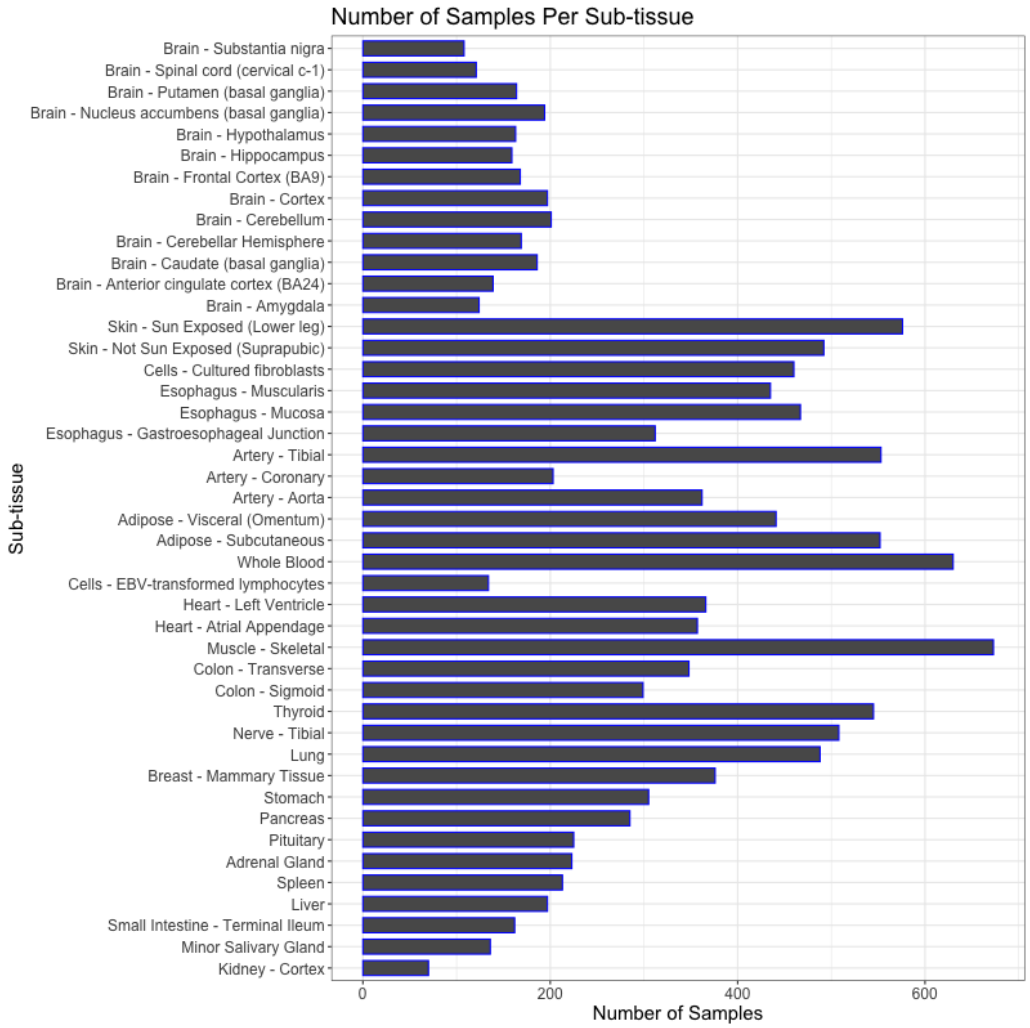


Figure 4.1: **Sample number per GTEx tissue subtype** used for network inference. (From Hu et al. [2022])

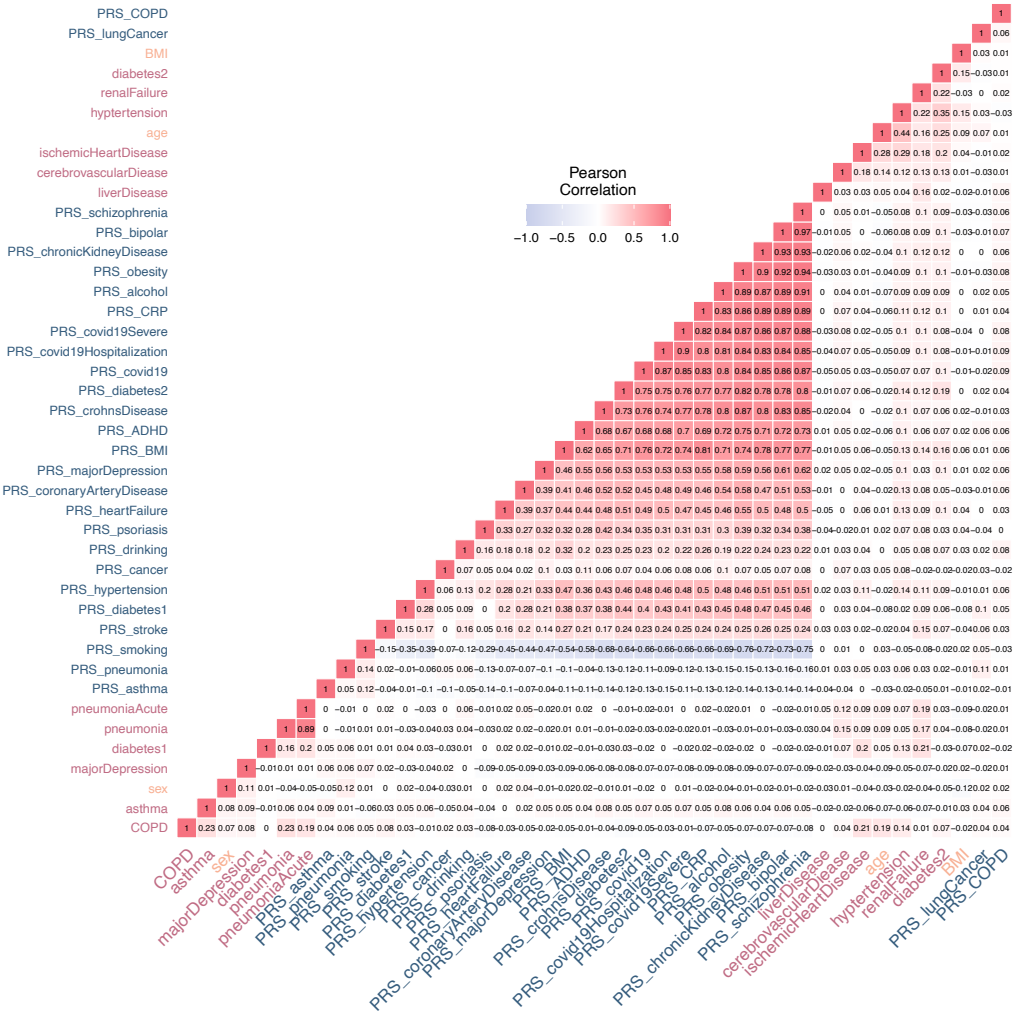


Figure 4.2: **Correlation of different modalities** derived from Polygenic Risk Scores (PRS) (blue), comorbidities (purple) and phenotypes (orange), given as pearson correlation. (From Hu et al. [2022])

Table 4.1: GWAS origin for PRS calculation. (From Hu et al. [2022])

Risk factor symptom	GWAS source	X samples in GWAS	X SNPs at 5e 08	disease class	Highly polygenic
ADHD	PGC: ADHD European June 2017 https://www.med.unc.edu/pgc/download-results/adhd/	53293	317	Psychiatric / Neurological	1
Alcohol Dependence	PGC: Alcohol Dependence 2018 pgc_alcdep.eur_unrel_genotyped.aug2018_release.txt.gz https://www.med.unc.edu/pgc/download-results/alcohol-dependence/?choice=AlcoholAlcohol+Dependence+%28ALCDEP%29	28757	4	Urinary system	0
Asthma	GWAS catalog: https://www.ebi.ac.uk/gwas/studies/GCST008916	394283	5286	Respiratory system	2
BIP	Inhouse data: Inistitute of Computational Health - Helmholtz Zentrum Munich server	147172	240	Psychiatric / Neurological	1
BMI Giant	Giant Consortium: 2018 bmi.giant-ukbb.meta-analysis.combined.23May2018.HapMap2_only.txt.gz https://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files	806834	42621	Body measurements	2
Cancer	UKBioBank: Cancer diagnosed by doctor https://www.dropbox.com/s/3rzc6gffq1nve4u/2453.gwas.imputed_v3.both_sexes.tsv.bgz?dl=0	359981	214	Cancer	2
Chronic kidney disease	GWAS catalog: https://www.ebi.ac.uk/gwas/studies/GCST008064 ->CKD_overall_ALL_JW_20180223_nstud30.dbgap.txt.gz	625219	1443	Urinary system	2
COPD	UKBioBank: Doctor diagnosed COPD https://www.dropbox.com/s/idukd82gkpew7zm/22137.gwas.imputed_v3.both_sexes.tsv.bgz?dl=0	50553	1156	Respiratory system	1
Coronary artery disease	Inhouse data: Max-Planck Institute of Psychiatry server	184305	2046	Cardiovascular	1
Crohns disease	GWAS catalog: https://www.ebi.ac.uk/gwas/studies/GCST004132 ->28067908-GCST004132-EFO_0000384.h.tsv.gz	40266	11016	Gastrointestinal	1
CRP	Inhouse data: Max-Planck Institute of Psychiatry server	204402	7201	Immune Response	1
Diabetes Type 1	UKBioBank: Type 1 diabetes - both sexes https://www.dropbox.com/s/qnskzu1obrpc5t6/E4_DM1.gwas.imputed_v3.both_sexes.tsv.bgz?dl=0	361194	8596	Digestive system	2
Diabetes Type 2	GWAS catalog: https://www.ebi.ac.uk/gwas/publications/30054458 ->30054458-GCST006867-EFO_0001360-build37.f.tsv.gz	659316	5984	Digestive system	2
Drinking	UKBioBank: Alcohol intake frequency https://www.dropbox.com/s/bi7t4rekkhpa4ks/1558.gwas.imputed_v3.both_sexes.tsv.bgz?dl=0	360726	8384	Urinary system	2
Hypertension	UKBioBank: Hypertension https://www.dropbox.com/s/p1yvjecpfg50vw8/I9_HYPTENS.gwas.imputed_v3.both_sexes.tsv.bgz?dl=0	361194	803	Cardiovascular	2
Lung cancer	UKBioBank: Lung cancer and mesothelioma https://www.dropbox.com/s/qk8gnbi9qmvo4w4/LUNG_CANCER_MESOT.gwas.imputed_v3.both_sexes.tsv.bgz?dl=0	361194	652	Respiratory system	2
MDD	PGC: 2019 PGC UKB Depression Genome-Wide https://www.med.unc.edu/pgc/download-results/mdd/	500199	4625	Psychiatric / Neurological	2
Pneumonia	UKBioBank: Pneumonias (Asthma/COPD comorbidities) https://www.dropbox.com/s/154y9toou0xuvbi/PNEUMONIA.gwas.imputed_v3.both_sexes.tsv.bgz?dl=0	361194	574	Respiratory system	2
SCZ	PGC: SCZ2 SNP results https://www.med.unc.edu/pgc/download-results/scz/	150064	11799	Psychiatric / Neurological	1
Smoking status	Inhouse data: Max-Planck Institute of Psychiatry server	74053	129	Respiratory system	1
Stroke	UKBioBank: Ischaemic stroke https://www.dropbox.com/s/e8qvytb057weseo/I9_STR_EXH.gwas.imputed_v3.both_sexes.tsv.bgz?dl=0	361194	704	Cardiovascular	2
Obesity	GWAS catalog: https://www.ebi.ac.uk/gwas/studies/GCST007241	7916	180	Body measurements	1
Heart Failure	GWAS catalog: https://www.ebi.ac.uk/gwas/studies/GCST009541	977323	295	Cardiovascular disease	2
Psoriasis	GWAS catalog: https://www.ebi.ac.uk/gwas/studies/GCST005527	33394	2189	Immunesystem disease	1
Covid19hg	The covid19 Host Genetics Initiative: https://www.covid19hg.org/results/r6/	2586691	703	Covid19	2
Covid19hghospitalized	The covid19 Host Genetics Initiative: https://www.covid19hg.org/results/r6/	2085803	3616	Covid19	2
Covid19hgsevere	The covid19 Host Genetics Initiative: https://www.covid19hg.org/results/r6/	1010654	1231	Covid19	2

Bibliography

K. Abbas, A. Abbasi, S. Dong, L. Niu, L. Yu, B. Chen, S.-M. Cai, and Q. Hasan. Application of network link prediction in drug discovery. *BMC Bioinformatics*, 22(1):187, Apr. 2021. ISSN 1471-2105. doi: 10.1186/s12859-021-04082-y.

H. Abelaira, G. Réus, M. Neotti, and J. Quevedo. The role of mTOR in depression and antidepressant responses. *Life Sci*, 101:10–14, 2014.

P. Afshar, K. N. Plataniotis, and A. Mohammadi. Capsule Networks for Brain Tumor Classification based on MRI Images and Course Tumor Boundaries, Nov. 2018. arXiv:1811.00597 [cs].

A. Ahmed, N. Shervashidze, S. Narayanamurthy, V. Josifovski, and A. J. Smola. Distributed large-scale natural graph factorization. In *Proceedings of the 22nd international conference on World Wide Web*, pages 37–48, 2013.

R. Albert, H. Jeong, and A.-L. Barabási. Diameter of the world-wide web. *nature*, 401(6749): 130–131, 1999. Publisher: Nature Publishing Group UK London.

S. A. Alcalá-Corona, S. Sandoval-Motta, J. Espinal-Enríquez, and E. Hernández-Lemus. Modularity in Biological Networks. *Frontiers in Genetics*, 12, 2021. ISSN 1664-8021. doi: 10.3389/fgene.2021.701331.

S. A. Aleksander, J. Balhoff, S. Carbon, J. M. Cherry, H. J. Drabkin, D. Ebert, M. Feuermann, P. Gaudet, N. L. Harris, and others. The gene ontology knowledgebase in 2023. *Genetics*, 224 (1):iyad031, 2023. Publisher: Oxford University Press US.

G. Alexanderson. About the cover: Euler and Königsberg's Bridges: A historical view. *Bulletin of the american mathematical society*, 43(4):567–573, 2006.

A. Amara, M. A. Hadj Taieb, and M. Ben Aouicha. Network representation learning systematic review: Ancestors and current development state. *Machine Learning with Applications*, 6: 100130, Dec. 2021. ISSN 2666-8270. doi: 10.1016/j.mlwa.2021.100130.

K. M. Anderson, M. A. Collins, R. Kong, K. Fang, J. Li, T. He, A. M. Chekroud, B. T. T. Yeo, and A. J. Holmes. Convergent molecular, cellular, and cortical neuroimaging signatures of major depressive disorder. *Proceedings of the National Academy of Sciences*, 117(40):25138–25149, Oct. 2020. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2008004117.

C. G. Ardanaz, M. J. Ramírez, and M. Solas. Brain Metabolic Alterations in Alzheimer's Disease. *International Journal of Molecular Sciences*, 23(7):3785, Mar. 2022. ISSN 1422-0067. doi: 10.3390/ijms23073785.

R. Argelaguet, B. Velten, D. Arnol, S. Dietrich, T. Zenz, J. C. Marioni, F. Buettner, W. Huber, and O. Stegle. Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, 14(6), June 2018. ISSN 1744-4292, 1744-4292. doi: 10.15252/msb.20178124.

J. Arloth, D. M. Bader, S. Röh, and A. Altmann. Re-Annotator: Annotation Pipeline for Microarray Probe Sequences. *PLOS ONE*, 10(10):e0139516, Oct. 2015a. ISSN 1932-6203. doi: 10.1371/journal.pone.0139516.

J. Arloth, R. Bogdan, P. Weber, G. Frishman, A. Menke, K. V. Wagner, G. Balsevich, M. V. Schmidt, N. Karbalai, D. Czamara, A. Altmann, D. Trümbach, W. Wurst, D. Mehta, M. Uhr, T. Klengel, A. Erhardt, C. E. Carey, E. D. Conley, Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium (PGC), A. Ruepp, B. Müller-Myhsok, A. R. Hariri, E. B. Binder, and Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium PGC. Genetic Differences in the Immediate Transcriptome Response to Stress Predict Risk-Related Brain Function and Psychiatric Disorders. *Neuron*, 86(5):1189–1202, June 2015b. ISSN 1097-4199. doi: 10.1016/j.neuron.2015.05.034.

M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, and others. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000. Publisher: Nature Publishing Group.

M. Ashtiani, A. Salehzadeh-Yazdi, Z. Razaghi-Moghadam, H. Hennig, O. Wolkenhauer, M. Mirzaie, and M. Jafari. A systematic survey of centrality measures for protein-protein interaction networks. *BMC Systems Biology*, 12(1):80, July 2018. ISSN 1752-0509. doi: 10.1186/s12918-018-0598-2.

R. E. Baker, J.-M. Peña, J. Jayamohan, and A. Jérusalem. Mechanistic models versus machine learning, a fight worth fighting for the biological community? *Biology Letters*, 14(5):20170660, May 2018. doi: 10.1098/rsbl.2017.0660. Publisher: Royal Society.

D. Bano, D. Ehninger, and G. Bagetta. Decoding metabolic signatures in Alzheimer's disease: a mitochondrial perspective. *Cell Death Discovery*, 9(1):1–4, Dec. 2023. ISSN 2058-7716. doi: 10.1038/s41420-023-01732-3. Publisher: Nature Publishing Group.

A.-L. Barabasi and Z. Oltvai. Network Biology: Understanding The Cell's Functional Organization. *Nature reviews. Genetics*, 5:101–13, Mar. 2004. doi: 10.1038/nrg1272.

A.-L. Barabási and M. Pósfai. *Network science*. Cambridge University Press, Cambridge, 2016. ISBN 978-1-107-07626-6 1-107-07626-9.

A.-L. Barabási, N. Gulbahce, and J. Loscalzo. Network medicine: a network-based approach to human disease. *Nature reviews genetics*, 12(1):56–68, 2011. Publisher: Nature Publishing Group UK London.

J. S. Baras and G. Theodorakopoulos. Path problems in networks. *Synthesis Lectures on Communication Networks*, 3(1):1–77, 2010. Publisher: Morgan & Claypool Publishers.

R. Batra, M. Arnold, M. A. Wörheide, M. Allen, X. Wang, C. Blach, A. I. Levey, N. T. Seyfried, N. Ertekin-Taner, D. A. Bennett, and others. The landscape of metabolic brain alterations in Alzheimer's disease. *Alzheimer's & Dementia*, 19(3):980–998, 2023. Publisher: Wiley Online Library.

P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner, C. Gulcehre, F. Song, A. Ballard, J. Gilmer, G. Dahl, A. Vaswani, K. Allen, C. Nash, V. Langston, C. Dyer, N. Heess, D. Wierstra, P. Kohli, M. Botvinick, O. Vinyals, Y. Li, and R. Pascanu. Relational inductive biases, deep learning, and graph networks, Oct. 2018. arXiv:1806.01261 [cs, stat].

M. Belkin and P. Niyogi. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Computation*, 15(6):1373–1396, June 2003. ISSN 0899-7667, 1530-888X. doi: 10.1162/089976603321780317.

C. Bellenguez, F. Küçükali, I. E. Jansen, L. Kleineidam, S. Moreno-Grau, N. Amin, A. C. Naj, R. Campos-Martin, B. Grenier-Boley, V. Andrade, and others. New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nature genetics*, 54(4):412–436, 2022. Publisher: Nature Publishing Group US New York.

R. Bellman. On a routing problem. *Quarterly of applied mathematics*, 16(1):87–90, 1958.

B. Bielorai, T. Fisher, D. Waldman, Y. Lerenthal, A. Nissenkorn, T. Tohami, D. Marek, N. Amariglio, and A. Toren. Acute lymphoblastic leukemia in early childhood as the presenting sign of ataxia-telangiectasia variant. *Pediatric Hematology and Oncology*, 30(6):574–582, Sept. 2013. ISSN 1521-0669. doi: 10.3109/08880018.2013.777949.

V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, Oct. 2008. doi: 10.1088/1742-5468/2008/10/P10008.

R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Ré, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, and P. Liang. On the Opportunities and Risks of Foundation Models, July 2022. arXiv:2108.07258 [cs].

S. Bonner, I. P. Barrett, C. Ye, R. Swiers, O. Engkvist, A. Bender, C. T. Hoyt, and W. L. Hamilton. A review of biomedical datasets relating to drug discovery: a knowledge graph perspective. *Briefings in Bioinformatics*, 23(6):bbac404, Nov. 2022. ISSN 1477-4054. doi: 10.1093/bib/bbac404.

A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko. Translating Embeddings for Modeling Multi-relational Data. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

R. Bouali-Benazzouz and A. Benazzouz. Covid-19 Infection and Parkinsonism: Is There a Link? *Movement Disorders*, 36(8):1737–1743, Aug. 2021. ISSN 0885-3185, 1531-8257. doi: 10.1002/mds.28680.

P. C. Bradshaw. Acetyl-CoA Metabolism and Histone Acetylation in the Regulation of Aging and Lifespan. *Antioxidants*, 10(4):572, Apr. 2021. ISSN 2076-3921. doi: 10.3390/antiox10040572.

M. Breen. Gene expression in cord blood links genetic risk for neurodevelopmental disorders with maternal psychological distress and adverse childhood outcomes. *Brain Behav. Immun*, 73:320–330, 2018.

L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, Oct. 2001. ISSN 1573-0565. doi: 10.1023/A:1010933404324.

A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. *Computer networks*, 33(1-6):309–320, 2000. Publisher: Elsevier.

A. S. Buchman, L. Yu, P. A. Boyle, J. A. Schneider, P. L. De Jager, and D. A. Bennett. Higher brain BDNF gene expression is associated with slower cognitive decline in older adults. *Neurology*, 86(8):735–741, 2016. Publisher: AAN Enterprises.

S. Budach and A. Marsico. pysster: classification of biological sequences by learning sequence and structure motifs with convolutional neural networks. *Bioinformatics*, 34(17):3035–3037, Sept. 2018. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/bty222.

S. Budd Haeberlein, P. Aisen, F. Barkhof, S. Chalkias, T. Chen, S. Cohen, G. Dent, O. Hansson, K. Harrison, C. Von Hehn, and others. Two randomized phase 3 studies of aducanumab in early Alzheimer's disease. *The journal of prevention of Alzheimer's disease*, 9(2):197–210, 2022. Publisher: Springer.

W. Burggren, K. Chapman, B. B. Keller, M. Monticino, and J. S. Torday. Interdisciplinarity in the Biological Sciences. In R. Frodeman, editor, *The Oxford Handbook of Interdisciplinarity*, page 0. Oxford University Press, Jan. 2017. ISBN 978-0-19-873352-2. doi: 10.1093/oxfordhb/9780198733522.013.9.

B. I. Bustos, K. Billingsley, C. Blauwendraat, J. R. Gibbs, Z. Gan-Or, D. Krainc, A. B. Singleton, S. J. Lubbe, and I. P. D. G. Consortium (IPDGC). Genome-wide contribution of common short-tandem repeats to Parkinson's disease genetic risk. *Brain*, 146(1):65–74, 2023. Publisher: Oxford University Press US.

K. Börner, S. Sanyal, A. Vespignani, and others. Network science. *Annu. rev. inf. sci. technol.*, 41(1):537–607, 2007.

Q. Cai, V. K. Mukku, and M. Ahmad. Coronary Artery Disease in Patients with Chronic Kidney Disease: A Clinical Update. *Current Cardiology Reviews*, 9(4):331–339, Nov. 2013. ISSN 1573-403X. doi: 10.2174/1573403X10666140214122234.

L. Cantini, E. Medico, S. Fortunato, and M. Caselle. Detection of gene communities in multi-networks reveals cancer drivers. *Scientific reports*, 5(1):17386, 2015. Publisher: Nature Publishing Group UK London.

S. Cao, W. Lu, and Q. Xu. Grarep: Learning graph representations with global structural information. In *Proceedings of the 24th ACM international on conference on information and knowledge management*, pages 891–900, 2015.

L. J. Carithers, K. Ardlie, M. Barcus, P. A. Branton, A. Britton, S. A. Buia, C. C. Compton, D. S. DeLuca, J. Peter-Demchok, E. T. Gelfand, P. Guan, G. E. Korzeniewski, N. C. Lockhart, C. A. Rabiner, A. K. Rao, K. L. Robinson, N. V. Roche, S. J. Sawyer, A. V. Segrè, C. E. Shive, A. M. Smith, L. H. Sobin, A. H. Undale, K. M. Valentino, J. Vaught, T. R. Young, H. M. Moore, and on behalf of the GTEx Consortium. A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. *Biopreservation and Biobanking*, 13(5): 311–319, Oct. 2015. ISSN 1947-5535, 1947-5543. doi: 10.1089/bio.2015.0032.

A. Cauchy and others. Méthode générale pour la résolution des systemes d'équations simultanées. *Comp. Rend. Sci. Paris*, 25(1847):536–538, 1847.

E. G. Cerami, B. E. Gross, E. Demir, I. Rodchenkov, O. Babur, N. Anwar, N. Schultz, G. D. Bader, and C. Sander. Pathway Commons, a web resource for biological pathway data. *Nucleic acids research*, 39(suppl_1):D685–D690, 2010. Publisher: Oxford University Press.

P. Chandak, K. Huang, and M. Zitnik. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10(1):67, 2023. Publisher: Nature Publishing Group UK London.

C. C. Chang, C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, 4(1):7, Dec. 2015. ISSN 2047-217X. doi: 10.1186/s13742-015-0047-8.

C.-Y. Chen and K.-M. Liao. Chronic Obstructive Pulmonary Disease is associated with risk of Chronic Kidney Disease: A Nationwide Case-Cohort Study. *Scientific Reports*, 6:25855, May 2016. ISSN 2045-2322. doi: 10.1038/srep25855.

L. Chen, Y.-H. Zhang, G. Huang, X. Pan, T. Huang, and Y.-D. Cai. Inferring novel genes related to oral cancer with a network embedding method and one-class learning algorithms. *Gene Therapy*, 26(12):465–478, Dec. 2019. ISSN 1476-5462. doi: 10.1038/s41434-019-0099-y. Number: 12 Publisher: Nature Publishing Group.

J. M. Chessells, G. Harrison, S. M. Richards, C. C. Bailey, F. G. Hill, B. E. Gibson, and I. M. Hann. Down's syndrome and acute lymphoblastic leukaemia: clinical features and response to treatment. *Archives of Disease in Childhood*, 85(4):321–325, Oct. 2001. ISSN 1468-2044. doi: 10.1136/adc.85.4.321.

B. G. Childs, M. Durik, D. J. Baker, and J. M. Van Deursen. Cellular senescence in aging and age-related disease: from mechanisms to therapy. *Nature medicine*, 21(12):1424–1435, 2015. Publisher: Nature Publishing Group.

H. Cho, B. Berger, and J. Peng. Compact Integration of Multi-Network Topology for Functional Analysis of Genes. *Cell Systems*, 3(6):540–548.e5, Dec. 2016. ISSN 24054712. doi: 10.1016/j.cels.2016.10.017.

M. Civelek and A. J. Lusis. Systems genetics approaches to understand complex traits. *Nature reviews. Genetics*, 15(1):34–48, Jan. 2014. ISSN 1471-0064 1471-0056. doi: 10.1038/nrg3575. Place: England.

B. S. Clarke and J. E. Mittenthal. Modularity and reliability in the organization of organisms. *Bulletin of Mathematical Biology*, 54(1):1–20, 1992. Publisher: Elsevier.

T. U. Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531, Nov. 2022. ISSN 0305-1048. doi: 10.1093/nar/gkac1052. _eprint: https://academic.oup.com/nar/article-pdf/51/D1/D523/48441158/gkac1052.pdf.

T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to algorithms*. MIT press, 2022.

G. Corso, L. Cavalleri, D. Beaini, P. Liò, and P. Veličković. Principal Neighbourhood Aggregation for Graph Nets, Dec. 2020. arXiv:2004.05718 [cs, stat].

B. M. Crossley, J. Bai, A. Glaser, R. Maes, E. Porter, M. L. Killian, T. Clement, and K. Tooheykurth. Guidelines for Sanger sequencing and molecular assay monitoring. *Journal of Veterinary Diagnostic Investigation*, 32(6):767–775, 2020. doi: 10.1177/1040638720905833. _eprint: https://doi.org/10.1177/1040638720905833.

J. Cummings, Y. Zhou, G. Lee, K. Zhong, J. Fonseca, and F. Cheng. Alzheimer's disease drug development pipeline: 2023. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 9(2):e12385, 2023. ISSN 2352-8737. doi: 10.1002/trc2.12385. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/trc2.12385.

R. S. Da Cruz, E. Chen, M. Smith, J. Bates, and S. De Assis. Diet and transgenerational epigenetic inheritance of breast cancer: the role of the paternal germline. *Frontiers in Nutrition*, 7:93, 2020. Publisher: Frontiers Media SA.

C. Darwin. *On the Origin of Species by Means of Natural Selection*. Murray, London, 1859.

R. Das, S. Dhuliawala, M. Zaheer, L. Vilnis, I. Durugkar, A. Krishnamurthy, A. Smola, and A. McCallum. Go for a Walk and Arrive at the Answer: Reasoning Over Paths in Knowledge Bases using Reinforcement Learning, Dec. 2018. arXiv:1711.05851 [cs].

C. P. Day. Who gets alcoholic liver disease: nature or nurture? *Journal of the Royal College of Physicians of London*, 34(6):557, 2000. Publisher: Royal College of Physicians.

W. J. Deardorff, E. Feen, and G. T. Grossberg. The use of cholinesterase inhibitors across all stages of Alzheimer's disease. *Drugs & aging*, 32:537–547, 2015. Publisher: Springer.

P. C. Deedwania. Metabolic syndrome and vascular disease: is nature or nurture leading the new epidemic of cardiovascular disease? *Circulation*, 109(1):2–4, 2004. Publisher: Am Heart Assoc.

V. Demichev. A time-resolved proteomic and prognostic map of COVID-19. *OPEN ACCESS*, page 23, 2021.

E. Demir, M. P. Cary, S. Paley, K. Fukuda, C. Lemer, I. Vastrik, G. Wu, P. D'eustachio, C. Schaefer, J. Luciano, and others. The BioPAX community standard for pathway data sharing. *Nature biotechnology*, 28(9):935–942, 2010. Publisher: Nature Publishing Group US New York.

M. E. Demir, Z. Ercan, E. Y. Karakas, T. Ulas, and H. Buyukhatipoglu. Crohnic Kidney Disease: Recurrent Acute Kidney Failure in a Patient With Crohn's Disease. *North American Journal of Medical Sciences*, 6(12):648–649, Dec. 2014. ISSN 2250-1541. doi: 10.4103/1947-2714.147983.

B. Di, H. Jia, O. J. Luo, F. Lin, K. Li, Y. Zhang, H. Wang, H. Liang, J. Fan, and Z. Yang. Identification and validation of predictive factors for progression to severe COVID-19 pneumonia by proteomics. *Signal Transduction and Targeted Therapy*, 5:217, Oct. 2020. ISSN 2095-9907. doi: 10.1038/s41392-020-00333-1.

Y. Dong, N. V. Chawla, and A. Swami. metapath2vec: Scalable Representation Learning for Heterogeneous Networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 135–144, Halifax NS Canada, Aug. 2017. ACM. ISBN 978-1-4503-4887-4. doi: 10.1145/3097983.3098036.

J. Duchi, E. Hazan, and Y. Singer. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of machine learning research*, 2011.

C. Ducruet and L. Beauguitte. Spatial Science and Network Science: Review and Outcomes of a Complex Relationship. *Networks and Spatial Economics*, 14(3):297–316, Dec. 2014. ISSN 1572-9427. doi: 10.1007/s11067-013-9222-6.

R. Duman, G. Aghajanian, G. Sanacora, and J. Krystal. Synaptic plasticity and depression: new insights from stress and rapid-acting antidepressants. *Nat. Med*, 22:238–249, 2016.

A. D'Alessandro, T. Thomas, M. Dzieciatkowska, R. C. Hill, R. O. Francis, K. E. Hudson, J. C. Zimring, E. A. Hod, S. L. Spitalnik, and K. C. Hansen. Serum Proteomics in COVID-19 Patients: Altered Coagulation and Complement Status as a Function of IL-6 Level. *Journal of Proteome Research*, 19(11):4417–4427, Nov. 2020. ISSN 1535-3893. doi: 10.1021/acs.jproteome.0c00365. Publisher: American Chemical Society.

S. F. Ehrlich, C. P. Quesenberry, Jr., S. K. Van Den Eeden, J. Shan, and A. Ferrara. Patients Diagnosed With Diabetes Are at Increased Risk for Asthma, Chronic Obstructive Pulmonary Disease, Pulmonary Fibrosis, and Pneumonia but Not Lung Cancer. *Diabetes Care*, 33(1): 55–60, Oct. 2009. ISSN 0149-5992. doi: 10.2337/dc09-0880.

L. Ehrlinger and W. Wöß. Towards a definition of knowledge graphs. *SEMANTiCS (Posters, Demos, SuCCESS)*, 48(1-4):2, 2016.

S. Elezkurtaj, S. Greuel, J. Ihlow, E. G. Michaelis, P. Bischoff, C. A. Kunze, B. V. Sinn, M. Gerhold, K. Hauptmann, B. Ingold-Heppner, F. Miller, H. Herbst, V. M. Corman, H. Martin, H. Radbruch, F. L. Heppner, and D. Horst. Causes of death and comorbidities in hospitalized patients with COVID-19. *Scientific Reports*, 11(1):4263, Feb. 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-82862-5. Number: 1 Publisher: Nature Publishing Group.

E. Engle. Human genetic disorders of axon guidance. Cold Spring Harb. *Perspect. Biol*, 2:001784–001784, 2010.

L. Euler. Solutio problematis ad geometriam situs pertinentis. *Commentarii academiae scientiarum Petropolitanae*, pages 128–140, 1741.

M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. *ACM SIGCOMM computer communication review*, 29(4):251–262, 1999. Publisher: ACM New York, NY, USA.

J. Fan, A. Cannistra, I. Fried, T. Lim, T. Schaffner, M. Crovella, B. Hescott, and M. D. Leiserson. A Multi-Species Functional Embedding Integrating Sequence and Network Structure, Dec. 2017a.

X. Fan, J. Liang, Z. Wu, X. Shan, H. Qiao, and T. Jiang. Expression of HLA-DR genes in gliomas: correlation with clinicopathological features and prognosis. *Chinese Neurosurgical Journal*, 3(1):27, Dec. 2017b. ISSN 2057-4967. doi: 10.1186/s41016-017-0090-7.

R. Feil and M. F. Fraga. Epigenetics and the environment: emerging patterns and implications. *Nature reviews genetics*, 13(2):97–109, 2012. Publisher: Nature Publishing Group UK London.

S. Felix Krueger. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, 27:1571, 2011.

D. A. Fell and A. Wagner. The small world of metabolism. *Nature biotechnology*, 18(11):1121–1122, 2000. Publisher: Nature Publishing Group.

S. Fishilevich, S. Zimmerman, A. Kohn, T. Iny Stein, T. Olender, E. Kolker, M. Safran, and D. Lancet. Genic insights from integrated human proteomics in GeneCards. *Database: The Journal of Biological Databases and Curation*, 2016:baw030, Mar. 2016. ISSN 1758-0463. doi: 10.1093/database/baw030.

M. H. Fitz-James and G. Cavalli. Molecular mechanisms of transgenerational epigenetic inheritance. *Nature Reviews Genetics*, 23(6):325–341, 2022. Publisher: Nature Publishing Group UK London.

G. Fond, K. Nemani, D. Etchecopar-Etchart, A. Loundou, D. C. Goff, S. W. Lee, C. Lancon, P. Auquier, K. Baumstarck, P.-M. Llorca, D. K. Yon, and L. Boyer. Association Between Mental Health Disorders and Mortality Among Patients With COVID-19 in 7 Countries: A Systematic Review and Meta-analysis. *JAMA Psychiatry*, 78(11):1208–1217, Nov. 2021. ISSN 2168-622X. doi: 10.1001/jamapsychiatry.2021.2274.

L. R. Ford. Network flow theory. *Academic Press*, 1956. Publisher: Rand Corporation Santa Monica, CA.

D. T. Forster, C. Boone, G. D. Bader, and B. Wang. BIONIC: Biological Network Integration using Convolutions. preprint, Bioinformatics, Mar. 2021.

S. Frara, A. Allora, L. Castellino, L. di Filippo, P. Loli, and A. Giustina. COVID-19 and the pituitary. *Pituitary*, pages 1–17, May 2021. ISSN 1386-341X. doi: 10.1007/s11102-021-01148-1.

L. C. Freeman. Centrality in social networks conceptual clarification. *Social Networks*, 1(3): 215–239, 1978. ISSN 0378-8733. doi: https://doi.org/10.1016/0378-8733(78)90021-7.

K. P. F.R.S. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901. doi: 10.1080/14786440109462720. Publisher: Taylor & Francis.

M. Färber, F. Bartscherer, C. Menne, and A. Rettinger. Linked data quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. *Semantic Web*, 9(1):77–129, Nov. 2017. ISSN 22104968, 15700844. doi: 10.3233/SW-170275.

M. Galkin, X. Yuan, H. Mostafa, J. Tang, and Z. Zhu. Towards Foundation Models for Knowledge Graph Reasoning, Apr. 2024. arXiv:2310.04562 [cs].

L. Galárraga, C. Teflioudi, K. Hose, and F. M. Suchanek. AMIE: Association rule mining under incomplete evidence in ontological knowledge bases. In *Proceedings of the 22nd International World Wide Web Conference, WWW'13*, pages 413–422, United States, 2013. Association for Computing Machinery. ISBN 978-1-4503-2035-1.

N. Gassen and T. Rein. Is there a role of autophagy in depression and antidepressant action? *Front. Psychiatry*, 10:337, 2019.

T. Ge, C.-Y. Chen, Y. Ni, Y.-C. A. Feng, and J. W. Smoller. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nature Communications*, 10(1):1776, Dec. 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-09718-5.

N. Gekakis, D. Staknis, H. B. Nguyen, F. C. Davis, L. D. Wilsbacher, D. P. King, J. S. Takahashi, and C. J. Weitz. Role of the CLOCK protein in the mammalian circadian mechanism. *Science*, 280(5369):1564–1569, 1998. Publisher: American Association for the Advancement of Science.

A.-V. Gendrel and E. Heard. Noncoding RNAs and epigenetic mechanisms during X-chromosome inactivation. *Annual review of cell and developmental biology*, 30:561–580, 2014. Publisher: Annual Reviews.

J. German. Bloom's syndrome. XX. The first 100 cancers. *Cancer Genetics and Cytogenetics*, 93(1):100–106, Jan. 1997. ISSN 0165-4608. doi: 10.1016/s0165-4608(96)00336-6.

P. E. Geyer, F. M. Arend, S. Doll, M.-L. Louiset, S. Virreira, J. B. Mu, M. Bruegel, M. T. Strauss, L. M. Holdt, M. Mann, and D. Teupser. High-resolution serum proteome trajectories in COVID-19 reveal patient-specific seroconversion. *EMBO Molecular Medicine*, page 16, 2021.

S. D. Ghiassian, J. Menche, and A.-L. Barabási. 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(4):e1004120, Aug. 2015. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1004120. Publisher: Public Library of Science.

J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl. Neural Message Passing for Quantum Chemistry. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1263–1272. PMLR, July 2017. ISSN: 2640-3498.

M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002. doi: 10.1073/pnas.122653799. _eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.122653799.

X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, Mar. 2010. ISSN: 1938-7228.

X. Glorot, A. Bordes, and Y. Bengio. Deep Sparse Rectifier Neural Networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 315–323. JMLR Workshop and Conference Proceedings, June 2011. ISSN: 1938-7228.

K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabási. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690, 2007. Publisher: National Acad Sciences.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

S. Goodwin, J. D. McPherson, and W. R. McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333–351, 2016. Publisher: Nature Publishing Group.

D. E. Gordon, J. Hiatt, M. Bouhaddou, V. V. Rezelj, S. Ulferts, H. Braberg, A. S. Jureka, K. Obernier, J. Z. Guo, J. Batra, R. M. Kaake, A. R. Weckstein, T. W. Owens, M. Gupta, S. Pourmal, E. W. Titus, M. Cakir, M. Soucheray, M. McGregor, Z. Cakir, G. Jang, M. J. O'Meara, T. A. Tummino, Z. Zhang, H. Foussard, A. Rojc, Y. Zhou, D. Kuchenov, R. Hüttenhain, J. Xu, M. Eckhardt, D. L. Swaney, J. M. Fabius, M. Ummadi, B. Tutuncuoglu, U. Rathore, M. Modak, P. Haas, K. M. Haas, Z. Z. C. Naing, E. H. Pulido, Y. Shi, I. Barrio-Hernandez, D. Memon, E. Petsalaki, A. Dunham, M. C. Marrero, D. Burke, C. Koh, T. Vallet, J. A. Silvas, C. M. Azumaya, C. Billesbølle, A. F. Brilot, M. G. Campbell, A. Diallo, M. S. Dickinson, D. Diwanji, N. Herrera, N. Hoppe, H. T. Kratochvil, Y. Liu, G. E. Merz, M. Moritz, H. C. Nguyen, C. Nowotny, C. Puchades, A. N. Rizo, U. Schulze-Gahmen, A. M. Smith, M. Sun, I. D. Young, J. Zhao, D. Asarnow, J. Biel, A. Bowen, J. R. Braxton, J. Chen, C. M. Chio, U. S. Chio, I. Deshpande, L. Doan, B. Faust, S. Flores, M. Jin, K. Kim, V. L. Lam, F. Li, J. Li, Y.-L. Li, Y. Li, X. Liu, M. Lo, K. E. Lopez, A. A. Melo, F. R. Moss, P. Nguyen, J. Paulino, K. I. Pawar, J. K. Peters, T. H. Pospiech, M. Safari, S. Sangwan, K. Schaefer, P. V. Thomas, A. C. Thwin, R. Trenker, E. Tse, T. K. M. Tsui, F. Wang, N. Whitis, Z. Yu, K. Zhang, Y. Zhang, F. Zhou, D. Saltzberg, QCRG Structural Biology Consortium, A. J. Hodder, A. S. Shun-Shion, D. M. Williams, K. M. White, R. Rosales, T. Kehrer, L. Miorin, E. Moreno, A. H. Patel, S. Rihn, M. M. Khalid, A. Vallejo-Gracia, P. Fozouni, C. R. Simoneau, T. L. Roth, D. Wu, M. A. Karim, M. Ghoussaini, I. Dunham, F. Berardi, S. Weigang, M. Chazal, J. Park, J. Logue, M. McGrath, S. Weston, R. Haupt, C. J. Hastie, M. Elliott, F. Brown, K. A. Burness, E. Reid, M. Dorward, C. Johnson, S. G. Wilkinson, A. Geyer, D. M. Giesel, C. Baillie, S. Raggett, H. Leech, R. Toth, N. Goodman, K. C. Keough, A. L. Lind, Zoonomia Consortium, R. J. Klesh, K. R. Hemphill, J. Carlson-Stevermer, J. Oki, K. Holden, T. Maures, K. S. Pollard, A. Sali, D. A. Agard, Y. Cheng, J. S. Fraser, A. Frost, N. Jura, T. Kortemme, A. Manglik, D. R. Southworth, R. M. Stroud, D. R. Alessi, P. Davies, M. B. Frieman, T. Ideker, C. Abate, N. Jouvenet, G. Kochs, B. Shoichet, M. Ott, M. Palmarini, K. M. Shokat, A. García-Sastre, J. A. Rassen, R. Grosse, O. S. Rosenberg, K. A. Verba, C. F. Basler, M. Vignuzzi, A. A. Peden, P. Beltrao, and N. J. Krogan. Comparative host-coronavirus protein interaction networks reveal pan-viral disease mechanisms. *Science*, 370(6521):eabe9403, Dec. 2020. doi: 10.1126/science.abe9403. Publisher: American Association for the Advancement of Science.

J. G. Greener, S. M. Kandathil, L. Moffat, and D. T. Jones. A guide to machine learning for biologists. *Nature Reviews Molecular Cell Biology*, 23(1):40–55, Jan. 2022. ISSN 1471-0080. doi: 10.1038/s41580-021-00407-0. Number: 1 Publisher: Nature Publishing Group.

E. A. Griffin, D. Staknis, and C. J. Weitz. Light-Independent Role of CRY1 and CRY2 in the Mammalian Circadian Clock. *Science*, 286(5440):768–771, 1999. doi: 10.1126/science.286.5440.768. _eprint: https://www.science.org/doi/pdf/10.1126/science.286.5440.768.

S. C.-. G. Group. Genomewide association study of severe Covid-19 with respiratory failure. *New England Journal of Medicine*, 383(16):1522–1534, 2020. Publisher: Mass Medical Soc.

A. Grover and J. Leskovec. node2vec: Scalable Feature Learning for Networks. *KDD : proceedings.*

International Conference on Knowledge Discovery & Data Mining, 2016:855–864, Aug. 2016. ISSN 2154-817X. doi: 10.1145/2939672.2939754.

L. Guarrera, M. Kurosaki, S.-K. Garattini, M. Gianni', G. Fasola, L. Rossit, M. Prisciandaro, M. Di Bartolomeo, M. Bolis, P. Rizzo, C. Nastasi, M. Foglia, A. Zanetti, G. Paroni, M. Terao, and E. Garattini. Anti-tumor activity of all-trans retinoic acid in gastric-cancer: gene-networks and molecular mechanisms. *Journal of Experimental & Clinical Cancer Research*, 42(1):298, Nov. 2023. ISSN 1756-9966. doi: 10.1186/s13046-023-02869-w.

A. Gupta, M. V. Madhavan, K. Sehgal, N. Nair, S. Mahajan, T. S. Sehrawat, B. Bikdeli, N. Ahluwalia, J. C. Ausiello, E. Y. Wan, D. E. Freedberg, A. J. Kirtane, S. A. Parikh, M. S. Maurer, A. S. Nordvig, D. Accili, J. M. Bathon, S. Mohan, K. A. Bauer, M. B. Leon, H. M. Krumholz, N. Uriel, M. R. Mehra, M. S. V. Elkind, G. W. Stone, A. Schwartz, D. D. Ho, J. P. Bilezikian, and D. W. Landry. Extrapulmonary manifestations of COVID-19. *Nature Medicine*, 26(7):1017–1032, July 2020. ISSN 1078-8956, 1546-170X. doi: 10.1038/s41591-020-0968-3.

M. N. Haidar, M. B. Islam, U. N. Chowdhury, M. R. Rahman, F. Huq, J. M. Quinn, and M. A. Moni. Network-based computational approach to identify genetic links between cardiomyopathy and its risk factors. *IET Systems Biology*, 14(2):75–84, Apr. 2020. ISSN 1751-8849, 1751-8857. doi: 10.1049/iet-syb.2019.0074.

E. Hall, J. Jönsson, J. K. Ofori, P. Volkov, A. Perfilyev, M. Dekker Nitert, L. Eliasson, C. Ling, and K. Bacos. Glucolipotoxicity Alters Insulin Secretion via Epigenetic Changes in Human Islets. *Diabetes*, 68(10):1965–1974, Oct. 2019. ISSN 0012-1797, 1939-327X. doi: 10.2337/db18-0900.

W. Hamilton, Z. Ying, and J. Leskovec. Inductive Representation Learning on Large Graphs. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017a.

W. L. Hamilton. *Graph Representation Learning*. Morgan \& Claypool Publishers, 2020.

W. L. Hamilton, R. Ying, and J. Leskovec. Representation learning on graphs: Methods and applications. *arXiv preprint arXiv:1709.05584*, 2017b.

M. S. Handcock, A. E. Raftery, and J. M. Tantrum. Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2):301–354, 2007. doi: https://doi.org/10.1111/j.1467-985X.2007.00471.x. _eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-985X.2007.00471.x.

S. Hapak, C. Rothlin, and S. Ghosh. aPKC in neuronal differentiation, maturation and function. *Neuronal Signal*, 3:20190019, 2019.

Y. Hasin, M. Seldin, and A. Lusis. Multi-omics approaches to disease. *Genome biology*, 18:1–15, 2017. Publisher: Springer.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer series in statistics. Springer, 2009. ISBN 978-0-387-84884-6.

K. He, X. Zhang, S. Ren, and J. Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015. doi: 10.1109/ICCV.2015.123.

J. D. Henao, M. Lauber, M. Azevedo, A. Grekova, M. List, C. Ogris, and B. Schubert. Multi-Omics Regulatory Network Inference in the Presence of Missing Data. preprint, Systems Biology, Apr. 2022.

D. Hendrycks and K. Gimpel. Gaussian Error Linear Units (GELUs), 2016. arXiv:1606.08415 [cs].

F. L. Heppner, R. M. Ransohoff, and B. Becher. Immune attack: the role of inflammation in Alzheimer disease. *Nature Reviews Neuroscience*, 16(6):358–372, 2015. Publisher: Nature Publishing Group UK London.

A. Hintze and C. Adami. Evolution of Complex Modular Biological Networks. *PLOS Computational Biology*, 4(2):1–12, Feb. 2008. doi: 10.1371/journal.pcbi.0040023. Publisher: Public Library of Science.

J. Ho, A. Jain, and P. Abbeel. Denoising Diffusion Probabilistic Models, Dec. 2020. arXiv:2006.11239 [cs, stat].

S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9 (8):1735–1780, Nov. 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. _eprint: https://direct.mit.edu/neco/article-pdf/9/8/1735/813796/neco.1997.9.8.1735.pdf.

A. E. Hoerl and R. W. Kennard. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1):55–67, 1970. ISSN 0040-1706. doi: 10.2307/1267351. Publisher: [Taylor & Francis, Ltd., American Statistical Association, American Society for Quality].

P. D. Hoff, A. E. Raftery, and M. S. Handcock. Latent Space Approaches to Social Network Analysis. *Journal of the American Statistical Association*, 97(460):1090–1098, Dec. 2002. ISSN 0162-1459, 1537-274X. doi: 10.1198/016214502388618906.

M. Horlacher, N. Wagner, L. Moyon, K. Kuret, N. Goedert, M. Salvatore, J. Ule, J. Gagneur, O. Winther, and A. Marsico. Towards in silico CLIP-seq: predicting protein-RNA interaction via sequence-to-signal learning. *Genome Biology*, 24(1):180, Aug. 2023. ISSN 1474-760X. doi: 10.1186/s13059-023-03015-7.

K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, Jan. 1989. ISSN 08936080. doi: 10.1016/0893-6080(89)90020-8.

Y. Hu, G. Rehawi, L. Moyon, N. Gerstner, C. Ogris, J. Knauer-Arloth, F. Bittner, A. Marsico, and N. S. Mueller. Network Embedding Across Multiple Tissues and Data Modalities Elucidates the Context of Host Factors Important for COVID-19 Infection. *Frontiers in Genetics*, 13:909714, 2022. ISSN 1664-8021. doi: 10.3389/fgene.2022.909714.

Y. Hu, S. Oleshko, S. Firmani, Z. Zhu, H. Cheng, M. Ulmer, M. Arnold, M. Colome-Tatche, J. Tang, S. Xhonneux, and A. Marsico. Path-based reasoning for biomedical knowledge graphs with BioKGC, Aug. 2024.

Y. Hua. Blocking endocytosis enhances short-term synaptic depression under conditions of normal availability of vesicles. *Neuron*, 80:343–349, 2013.

K. Huang, P. Chandak, Q. Wang, S. Havaldar, A. Vaid, J. Leskovec, G. Nadkarni, B. S. Glicksberg, N. Gehlenborg, and M. Zitnik. Zero-shot drug repurposing with geometric deep learning and clinician centered design, Mar. 2023.

F. Huguet, T. Leguay, E. Raffoux, P. Rousselot, N. Vey, A. Pigneux, N. Ifrah, and H. Dombret. Clofarabine for the treatment of adult acute lymphoid leukemia: the Group for Research on Adult Acute Lymphoblastic Leukemia intergroup. *Leukemia & Lymphoma*, 56(4):847–857, Apr. 2015. ISSN 1042-8194. doi: 10.3109/10428194.2014.887708. Publisher: Taylor & Francis.

E. L. Huttlin, R. J. Bruckner, J. A. Paulo, J. R. Cannon, L. Ting, K. Baltier, G. Colby, F. Gebreab, M. P. Gygi, H. Parzen, J. Szpyt, S. Tam, G. Zarraga, L. Pontano-Vaites, S. Swarup, A. E. White, D. K. Schweppe, R. Rad, B. K. Erickson, R. A. Obar, K. G. Guruharsha, K. Li, S. Artavanis-Tsakonas, S. P. Gygi, and J. W. Harper. Architecture of the human interactome defines protein communities and disease networks. *Nature*, 545(7655):505–509, May 2017. ISSN 1476-4687. doi: 10.1038/nature22366. Number: 7655 Publisher: Nature Publishing Group.

R. Ietswaart, B. M. Gyori, J. A. Bachman, P. K. Sorger, and L. S. Churchman. GeneWalk identifies relevant gene functions for a biological context using network representation learning. *Genome Biology*, 22(1):55, Dec. 2021. ISSN 1474-760X. doi: 10.1186/s13059-021-02264-8.

Z. Ignácio. New perspectives on the involvement of mTOR in depression as well as in the action of antidepressant drugs. *Br. J. Clin. Pharmacol*, 82:1280–1290, 2016.

B. P. Ingalls. *Mathematical modeling in systems biology: an introduction*. MIT press, 2013.

G. Iván and V. Grolmusz. When the Web meets the cell: using personalized PageRank for analyzing protein interaction networks. *Bioinformatics*, 27(3):405–407, Feb. 2011. ISSN 1367-4803. doi: 10.1093/bioinformatics/btq680.

P. Jaccard. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bull Soc Vaudoise Sci Nat*, 37:547–579, 1901.

M. Jackson, L. Marks, G. H. May, and J. B. Wilson. The genetic basis of disease. *Essays in biochemistry*, 62(5):643–723, 2018. Publisher: Portland Press Ltd.

A. Jaiswar, D. Arora, M. Malhotra, A. Shukla, and N. Rai. Broad Applications of Network Embeddings in Computational Biology, Genomics, Medicine, and Health. *Bioinformatics and Medical Applications: Big Data Using Deep Learning Algorithms*, pages 73–98, 2022. Publisher: Wiley Online Library.

M. Jamal, W. Van der Does, B. M. Elzinga, M. L. Molendijk, and B. W. Penninx. Association between smoking, nicotine dependence, and BDNF Val66Met polymorphism with BDNF concentrations in serum. *Nicotine & Tobacco Research*, 17(3):323–329, 2015. Publisher: Oxford University Press UK.

G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer Texts in Statistics. Springer New York, 2013. ISBN 978-1-4614-7138-7.

B. M. Javierre, A. F. Fernandez, J. Richter, F. Al-Shahrour, J. I. Martin-Subero, J. Rodriguez-Ubreva, M. Berdasco, M. F. Fraga, T. P. O'Hanlon, L. G. Rider, and others. Changes in the pattern of DNA methylation associate with twin discordance in systemic lupus erythematosus. *Genome research*, 20(2):170–179, 2010. Publisher: Cold Spring Harbor Lab.

H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, Oct. 2000. ISSN 0028-0836. doi: 10.1038/35036627.

G. Ji, S. He, L. Xu, K. Liu, and J. Zhao. Knowledge Graph Embedding via Dynamic Mapping Matrix. In C. Zong and M. Strube, editors, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 687–696, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1067.

I. T. Jolliffe. Principal Components in Regression Analysis. In *Principal Component Analysis*, pages 129–155. Springer New York, New York, NY, 1986. ISBN 978-1-4757-1904-8. doi: 10.1007/978-1-4757-1904-8_8.

L. L. Jones, D. A. McDonald, and P. R. Borum. Acylcarnitines: role in brain. *Progress in lipid research*, 49(1):61–75, 2010. Publisher: Elsevier.

J. Jurka, V. V. Kapitonov, O. Kohany, and M. V. Jurka. Repetitive sequences in complex genomes: structure and evolution. *Annu. Rev. Genomics Hum. Genet.*, 8:241–259, 2007. Publisher: Annual Reviews.

M. H. Kabir, R. Patrick, J. W. Ho, and M. D. O'Connor. Identification of active signaling pathways by integrating gene expression and protein interaction data. *BMC systems biology*, 12:77–87, 2018. Publisher: Springer.

M. Kaeberlein and V. Galvan. Rapamycin and Alzheimer's disease: time for a clinical trial? *Science translational medicine*, 11(476):eaar4289, 2019. Publisher: American Association for the Advancement of Science.

M. Kalligeros, F. Shehadeh, E. K. Mylona, G. Benitez, C. G. Beckwith, P. A. Chan, and E. Mylonakis. Association of obesity with disease severity among patients with coronavirus disease 2019. *Obesity*, 28(7):1200–1204, 2020. Publisher: Wiley Online Library.

A. Kamburov, C. Wierling, H. Lehrach, and R. Herwig. ConsensusPathDB—a database for integrating human functional interaction networks. *Nucleic Acids Research*, 37(Database issue): D623–D628, Jan. 2009. ISSN 0305-1048. doi: 10.1093/nar/gkn698.

A. Kamburov, K. Pentchev, H. Galicka, C. Wierling, H. Lehrach, and R. Herwig. ConsensusPathDB: toward a more complete picture of cell biology. *Nucleic Acids Research*, 39(suppl_1): D712–D717, Jan. 2011. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkq1156.

Z. A. Kaminsky, T. Tang, S.-C. Wang, C. Ptak, G. H. Oh, A. H. Wong, L. A. Feldcamp, C. Virtanen, J. Halfvarson, C. Tysk, and others. DNA methylation profiles in monozygotic and dizygotic twins. *Nature genetics*, 41(2):240–245, 2009. Publisher: Nature Publishing Group US New York.

R. Kandaswamy, A. McQuillin, D. Curtis, and H. Gurling. Tests of linkage and allelic association between markers in the 1p36 PRKCZ (protein kinase C zeta) gene region and bipolar affective disorder. *Am. J. Med. Genet. B Neuropsychiatr. Genet*, 159:201–209, 2012.

S. Kapur, A. G. Phillips, and T. R. Insel. Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? *Molecular Psychiatry*, 17(12):1174–1179, Dec. 2012. ISSN 1476-5578. doi: 10.1038/mp.2012.105.

G. Karlebach and R. Shamir. Modelling and analysis of gene regulatory networks. *Nature reviews Molecular cell biology*, 9(10):770–780, 2008. Publisher: Nature Publishing Group UK London.

L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953. Publisher: Springer.

D. Z. Kelley, E. L. Flam, E. Izumchenko, L. V. Danilova, H. A. Wulf, T. Guo, D. A. Singman, B. Afsari, A. M. Skaist, M. Considine, J. A. Welch, E. Stavrovskaya, J. A. Bishop, W. H. Westra, Z. Khan, W. M. Koch, D. Sidransky, S. J. Wheelan, J. A. Califano, A. V. Favorov, E. J. Fertig, and D. A. Gaykalova. Integrated Analysis of Whole-Genome ChIP-Seq and RNA-Seq Data of Primary Head and Neck Tumor Samples Associates HPV Integration Sites with Open Chromatin Marks. *Cancer Research*, 77(23):6538–6550, Nov. 2017. ISSN 0008-5472. doi: 10.1158/0008-5472.CAN-17-0833.

K. S. Kendler. What psychiatric genetics has taught us about the nature of psychiatric illness and what is left to learn. *Molecular Psychiatry*, 18(10):1058–1066, Oct. 2013. ISSN 1476-5578. doi: 10.1038/mp.2013.50.

K. S. Kendler, M. Gatz, C. O. Gardner, and N. L. Pedersen. A Swedish National Twin Study of Lifetime Major Depression. *American Journal of Psychiatry*, 163(1):109–114, 2006. doi: 10.1176/appi.ajp.163.1.109. _eprint: https://doi.org/10.1176/appi.ajp.163.1.109.

R. C. Kessler, P. Berglund, O. Demler, R. Jin, K. R. Merikangas, and E. E. Walters. Lifetime Prevalence and Age-of-Onset Distributions of DSM-IV Disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry*, 62(6): 593–602, June 2005. ISSN 0003-990X. doi: 10.1001/archpsyc.62.6.593. _eprint: https://jamanetwork.com/journals/jamapsychiatry/articlepdf/208678/yoa40305.pdf.

Z. J. Khitan, K.-V. Chin, K. Sodhi, M. Kheetan, A. Alsanani, and J. I. Shapiro. Gut Microbiome and Diet in Populations with Obesity: Role of the Na+/K+-ATPase Transporter Signaling in Severe COVID-19. *Obesity (Silver Spring, Md.)*, Jan. 2022. ISSN 1930-739X. doi: 10.1002/oby.23387.

H. I. Kim, C. R. Schultz, A. L. Buras, E. Friedman, A. Fedorko, L. Seamon, G. V. R. Chandramouli, G. L. Maxwell, A. S. Bachmann, and J. I. Risinger. Ornithine decarboxylase as a therapeutic target for endometrial cancer. *PLOS ONE*, 12(12):e0189044, Dec. 2017. ISSN 1932-6203. doi: 10.1371/journal.pone.0189044. Publisher: Public Library of Science.

S. Kim, E. Forno, R. Zhang, H. J. Park, Z. Xu, Q. Yan, N. Boutaoui, E. Acosta-Pérez, G. Canino, W. Chen, and J. C. Celedón. Expression Quantitative Trait Methylation Analysis Reveals Methylomic Associations With Gene Expression in Childhood Asthma. *Chest*, 158(5):1841–1856, Nov. 2020. ISSN 1931-3543. doi: 10.1016/j.chest.2020.05.601.

D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization, Jan. 2017. arXiv:1412.6980 [cs].

T. N. Kipf and M. Welling. Variational Graph Auto-Encoders, Nov. 2016. arXiv:1611.07308 [cs, stat].

T. N. Kipf and M. Welling. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv:1609.02907 [cs, stat]*, Feb. 2017. arXiv: 1609.02907.

H. Kitano. Systems biology: a brief overview. *science*, 295(5560):1662–1664, 2002. Publisher: American Association for the Advancement of Science.

P. Klaff, D. Riesner, and G. Steger. RNA structure and the regulation of gene expression. *Post-Transcriptional Control of Gene Expression in Plants*, pages 89–106, 1996. Publisher: Springer.

N. Klebanov. Genetic predisposition to infectious disease. *Cureus*, 10(8), 2018. Publisher: Cureus.

C. Knox, M. Wilson, C. M. Klinger, M. Franklin, E. Oler, A. Wilson, A. Pon, J. Cox, N. E. L. Chin, S. A. Strawbridge, M. Garcia-Patino, R. Kruger, A. Sivakumaran, S. Sanford, R. Doshi, N. Khetarpal, O. Fatokun, D. Doucet, A. Zubkowski, D. Y. Rayat, H. Jackson, K. Harford, A. Anjum, M. Zakir, F. Wang, S. Tian, B. Lee, J. Liigand, H. Peters, R. Q. R. Wang, T. Nguyen, D. So, M. Sharp, R. da Silva, C. Gabriel, J. Scantlebury, M. Jasinski, D. Ackerman, T. Jewison, T. Sajed, V. Gautam, and D. S. Wishart. DrugBank 6.0: the DrugBank Knowledgebase for 2024. *Nucleic Acids Research*, 52(D1):D1265–D1275, Jan. 2024. ISSN 1362-4962. doi: 10.1093/nar/gkad976.

D. Koschützki and F. Schreiber. Centrality Analysis Methods for Biological Networks and Their Application to Gene Regulatory Networks. *Gene Regulation and Systems Biology*, 2:193, 2008. doi: 10.4137/grsb.s702. Publisher: SAGE Publications.

D. Koschützki, H. Schwöbbermeyer, and F. Schreiber. Ranking of network elements based on functional substructures. *Journal of theoretical biology*, 248(3):471–479, 2007. Publisher: Elsevier.

P. Kowiański, G. Lietzau, E. Czuba, M. Waśkow, A. Steliga, and J. Moryś. BDNF: a key factor with multipotent impact on brain signaling and synaptic plasticity. *Cellular and molecular neurobiology*, 38:579–593, 2018. Publisher: Springer.

S. Krause, L. Hennig, A. Moro, D. Weissenborn, F. Xu, H. Uszkoreit, and R. Navigli. Sar-graphs: A language resource connecting linguistic knowledge with semantic relations from knowledge graphs. *Journal of Web Semantics*, 37:112–131, 2016. Publisher: Elsevier.

K. R. Kukurba and S. B. Montgomery. RNA sequencing and analysis. *Cold Spring Harbor Protocols*, 2015(11):pdb–top084970, 2015. Publisher: Cold Spring Harbor Laboratory Press.

A. Kumar, S. S. Singh, K. Singh, and B. Biswas. Link prediction techniques, applications, and performance: A survey. *Physica A: Statistical Mechanics and its Applications*, 553:124289, 2020. Publisher: Elsevier.

K. Kume, M. Zylka, S. Sriram, L. Shearman, D. Weaver, J.-J. Xu, E. Maywood, M. Hastings, and S. Reppert. mCRY1 and mCRY2 Are Essential Components of the Negative Limb of the Circadian Clock Feedback Loop. *Cell*, 98:193–205, Aug. 1999. doi: 10.1016/S0092-8674(00)81014-4.

H. T. Lam, M. L. Sbodio, M. M. Galindo, M. Zayats, R. Fernández-Díaz, V. Valls, G. Picco, C. B. Ramis, and V. López. Otter-Knowledge: benchmarks of multimodal knowledge graph representation learning from different sources for drug discovery, June 2023. arXiv:2306.12802 [cs, q-bio].

A. Lancichinetti, S. Fortunato, and J. Kertesz. Detecting the overlapping and hierarchical community structure of complex networks. *New Journal of Physics*, 11(3):033015, Mar. 2009. ISSN 1367-2630. doi: 10.1088/1367-2630/11/3/033015. arXiv:0802.1218 [cond-mat, physics:physics, stat].

H.-M. Lanoiselée, G. Nicolas, D. Wallon, A. Rovelet-Lecrux, M. Lacour, S. Rousseau, A.-C. Richard, F. Pasquier, A. Rollin-Sillaire, O. Martinaud, M. Quillard-Muraine, V. de la Sayette, C. Boutoleau-Bretonniere, F. Etcharry-Bouyx, V. Chauviré, M. Sarazin, I. le Ber, S. Epelbaum, T. Jonveaux, O. Rouaud, M. Ceccaldi, O. Félician, O. Godefroy, M. Formaglio, B. Croisile, S. Auriacombe, L. Chamard, J.-L. Vincent, M. Sauvée, C. Marelli-Tosi, A. Gabelle, C. Ozsancak, J. Pariente, C. Paquet, D. Hannequin, and D. Campion. APP, PSEN1, and PSEN2 mutations in early-onset Alzheimer disease: A genetic screening study of familial and sporadic cases. *PLoS Medicine*, 14(3):e1002270, Mar. 2017. ISSN 1549-1277. doi: 10.1371/journal.pmed.1002270.

V. Law, C. Knox, Y. Djoumbou, T. Jewison, A. C. Guo, Y. Liu, A. Maciejewski, D. Arndt, M. Wilson, V. Neveu, and others. DrugBank 4.0: shedding new light on drug metabolism. *Nucleic acids research*, 42(D1):D1091–D1097, 2014. Publisher: Oxford University Press.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

S. Lee, Y.-s. Lee, Y. Choi, A. Son, Y. Park, K.-M. Lee, J. Kim, J.-S. Kim, and V. N. Kim. The SARS-CoV-2 RNA Interactome. *Molecular Cell*, 81(13):2838–2850, 2021. Publisher: Elsevier.

T. I. Lee and R. A. Young. Transcriptional regulation and its misregulation in disease. *Cell*, 152 (6):1237–1251, 2013. Publisher: Elsevier.

M. Lehmann, K. Allers, C. Heldt, J. Meinhardt, F. Schmidt, Y. Rodriguez-Sillke, D. Kunkel, M. Schumann, C. Böttcher, C. Stahl-Hennig, S. Elezkurtaj, C. Bojarski, H. Radbruch, V. M. Corman, T. Schneider, C. Loddenkemper, V. Moos, C. Weidinger, A. A. Kühl, and B. Siegmund. Human small intestinal infection by SARS-CoV-2 is characterized by a mucosal infiltration with activated CD8+ T cells. *Mucosal Immunology*, 14(6):1381–1392, Nov. 2021. ISSN 1935-3456. doi: 10.1038/s41385-021-00437-z. Number: 6 Publisher: Nature Publishing Group.

T. G. Lewis and T. G. Lewis. *Network science: theory and practice*. Wiley, Hoboken, NJ, 2009. ISBN 978-0-470-33188-0.

G. Li, J. Luo, Q. Xiao, C. Liang, P. Ding, and B. Cao. Predicting MicroRNA-Disease Associations Using Network Topological Similarity Based on DeepWalk. *IEEE Access*, 5:24032–24039, 2017a. ISSN 2169-3536. doi: 10.1109/ACCESS.2017.2766758.

J. Li, H. Liu, S. P. Srivastava, Q. Hu, R. Gao, S. Li, M. Kitada, G. Wu, D. Koya, and K. Kanasaki. Endothelial FGFR1 (Fibroblast Growth Factor Receptor 1) Deficiency Contributes Differential Fibrogenic Effects in Kidney and Heart of Diabetic Mice. *Hypertension (Dallas, Tex.: 1979)*, 76(6):1935–1944, Dec. 2020. ISSN 1524-4563. doi: 10.1161/HYPERTENSIONAHA.120.15587.

M. M. Li, K. Huang, and M. Zitnik. Representation Learning for Networks in Biology and Medicine: Advancements, Challenges, and Opportunities. *arXiv:2104.04883 [cs, q-bio]*, Apr. 2021. arXiv: 2104.04883.

Y. Li and T. O. Tollefsbol. DNA methylation detection: bisulfite genomic sequencing analysis. *Epigenetics protocols*, pages 11–21, 2011. Publisher: Springer.

Y. Li, D. Tarlow, M. Brockschmidt, and R. Zemel. Gated Graph Sequence Neural Networks, Sept. 2017b. arXiv:1511.05493 [cs, stat].

D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *Proceedings of the twelfth international conference on Information and knowledge management*, pages 556–559, 2003.

B. Lima Giacobbo, J. Doorduin, H. C. Klein, R. A. J. O. Dierckx, E. Bromberg, and E. F. J. de Vries. Brain-Derived Neurotrophic Factor in Brain Disorders: Focus on Neuroinflammation. *Molecular Neurobiology*, 56(5):3295–3312, 2019. ISSN 0893-7648. doi: 10.1007/s12035-018-1283-6.

X.-Y. Lin, C.-D. He, T. Xiao, X. Jin, J. Chen, Y.-K. Wang, M. Liu, K.-B. Wang, Y. Jiang, H.-C. Wei, and H.-D. Chen. Acitretin induces apoptosis through CD95 signalling pathway in human cutaneous squamous cell carcinoma cell line SCL-1. *Journal of Cellular and Molecular Medicine*, 13(9a):2888–2898, Sept. 2009. ISSN 1582-1838. doi: 10.1111/j.1582-4934.2008.00397.x.

L. Liu, S.-Y. Ni, W. Yan, Q.-D. Lu, Y.-M. Zhao, Y.-Y. Xu, H. Mei, L. Shi, K. Yuan, Y. Han, J.-H. Deng, Y.-K. Sun, S.-Q. Meng, Z.-D. Jiang, N. Zeng, J.-Y. Que, Y.-B. Zheng, B.-N. Yang, Y.-M. Gong, A. V. Ravindran, T. Kosten, Y. K. Wing, X.-D. Tang, J.-L. Yuan, P. Wu, J. Shi, Y.-P. Bao, and L. Lu. Mental and neurological disorders and risk of COVID-19 susceptibility, illness severity and mortality: A systematic review, meta-analysis and call for action. *EClinicalMedicine*, 40:101111, Sept. 2021. ISSN 2589-5370. doi: 10.1016/j.eclinm.2021.101111.

S. Lobentanzer, P. Aloy, J. Baumbach, B. Bohar, V. J. Carey, P. Charoentong, K. Danhauser, T. Doğan, J. Dreo, I. Dunham, E. Farr, A. Fernandez-Torras, B. M. Gyori, M. Hartung, C. T. Hoyt, C. Klein, T. Korcsmaros, A. Maier, M. Mann, D. Ochoa, E. Pareja-Lorente, F. Popp, M. Preusse, N. Probul, B. Schwikowski, B. Sen, M. T. Strauss, D. Turei, E. Ulusoy, D. Waltemath, J. A. H. Wodke, and J. Saez-Rodriguez. Democratizing knowledge representation with BioCypher. *Nature Biotechnology*, June 2023. ISSN 1087-0156, 1546-1696. doi: 10.1038/s41587-023-01848-y.

J. Lonsdale, J. Thomas, M. Salvatore, R. Phillips, E. Lo, S. Shad, R. Hasz, G. Walters, F. Garcia, N. Young, B. Foster, M. Moser, E. Karasik, B. Gillard, K. Ramsey, S. Sullivan, J. Bridge, H. Magazine, J. Syron, J. Fleming, L. Siminoff, H. Traino, M. Mosavel, L. Barker, S. Jewell, D. Rohrer, D. Maxim, D. Filkins, P. Harbach, E. Cortadillo, B. Berghuis, L. Turner, E. Hudson, K. Feenstra, L. Sobin, J. Robb, P. Branton, G. Korzeniewski, C. Shive, D. Tabor, L. Qi, K. Groch, S. Nampally, S. Buia, A. Zimmerman, A. Smith, R. Burges, K. Robinson, K. Valentino, D. Bradbury, M. Cosentino, N. Diaz-Mayoral, M. Kennedy, T. Engel, P. Williams, K. Erickson, K. Ardlie, W. Winckler, G. Getz, D. DeLuca, D. MacArthur, M. Kellis, A. Thomson, T. Young, E. Gelfand, M. Donovan, Y. Meng, G. Grant, D. Mash, Y. Marcus, M. Basile, J. Liu, J. Zhu, Z. Tu, N. J. Cox, D. L. Nicolae, E. R. Gamazon, H. K. Im, A. Konkashbaev, J. Pritchard, M. Stevens, T. Flutre, X. Wen, E. T. Dermitzakis, T. Lappalainen, R. Guigo, J. Monlong, M. Sammeth, D. Koller, A. Battle, S. Mostafavi, M. McCarthy, M. Rivas, J. Maller, I. Rusyn, A. Nobel, F. Wright, A. Shabalin, M. Feolo, N. Sharopova, A. Sturcke, J. Paschal, J. M. Anderson, E. L. Wilder, L. K. Derr, E. D. Green, J. P. Struewing, G. Temple, S. Volpi,

J. T. Boyer, E. J. Thomson, M. S. Guyer, C. Ng, A. Abdallah, D. Colantuoni, T. R. Insel, S. E. Koester, A. R. Little, P. K. Bender, T. Lehner, Y. Yao, C. C. Compton, J. B. Vaught, S. Sawyer, N. C. Lockhart, J. Demchok, and H. F. Moore. The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, 45(6):580–585, June 2013. ISSN 1546-1718. doi: 10.1038/ng.2653. Number: 6 Publisher: Nature Publishing Group.

R. Loomes, L. Hull, and W. P. L. Mandy. What is the male-to-female ratio in autism spectrum disorder? A systematic review and meta-analysis. *Journal of the American Academy of Child & Adolescent Psychiatry*, 56(6):466–474, 2017. Publisher: Elsevier.

I. Loshchilov and F. Hutter. Decoupled Weight Decay Regularization, Jan. 2019. arXiv:1711.05101 [cs, math].

Y. Lu, Y. Guo, and A. Korhonen. Link prediction in drug-target interactions network using similarity indices. *BMC Bioinformatics*, 18:39, Jan. 2017. ISSN 1471-2105. doi: 10.1186/s12859-017-1460-z.

M. Lynch. Mutation and human exceptionalism: our future genetic load. *Genetics*, 202(3): 869–875, 2016. Publisher: Oxford University Press.

L. Lü and T. Zhou. Link prediction in complex networks: A survey. *Physica A: statistical mechanics and its applications*, 390(6):1150–1170, 2011. Publisher: Elsevier.

A. L. Maas, A. Y. Hannun, A. Y. Ng, and others. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3. Atlanta, GA, 2013. Issue: 1.

M. G. Machado, L. P. Tavares, G. V. S. Souza, C. M. Queiroz-Junior, F. R. Ascenção, M. E. Lopes, C. C. Garcia, G. B. Menezes, M. Perretti, R. C. Russo, M. M. Teixeira, and L. P. Sousa. The Annexin A1/FPR2 pathway controls the inflammatory response and bacterial dissemination in experimental pneumococcal pneumonia. *FASEB journal: official publication of the Federation of American Societies for Experimental Biology*, 34(2):2749–2764, Feb. 2020. ISSN 1530-6860. doi: 10.1096/fj.201902172R.

J. P. MacKeigan and D. A. Krueger. Differentiating the mTOR inhibitors everolimus and sirolimus in the treatment of tuberous sclerosis complex. *Neuro-oncology*, 17(12):1550–1559, 2015. Publisher: Society for Neuro-Oncology.

A. G. Malykh and M. R. Sadaie. Piracetam and piracetam-like drugs: from basic science to novel clinical applications to CNS disorders. *Drugs*, 70(3):287–312, Feb. 2010. Publisher: Springer Science and Business Media LLC.

J. B. Mannick and D. W. Lamming. Targeting the biology of aging with mTOR inhibitors. *Nature Aging*, pages 1–19, 2023. Publisher: Nature Publishing Group US New York.

J. Martens and R. Grosse. Optimizing Neural Networks with Kronecker-factored Approximate Curvature. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2408–2417, Lille, France, July 2015. PMLR.

N. Meinshausen and P. Buehlmann. Stability Selection, May 2009. arXiv:0809.2932 [stat].

A. Meissner. Epigenetic modifications in pluripotent and differentiated cells. *Nature biotechnology*, 28(10):1079–1088, 2010. Publisher: Nature Publishing Group US New York.

J. M. Mejia-Vilet, X. L. Zhang, C. Cruz, M. L. Cano-Verduzco, J. P. Shapiro, H. N. Nagaraja, L. E. Morales-Buenrostro, and B. H. Rovin. Urinary Soluble CD163: a Novel Noninvasive Biomarker of Activity for Lupus Nephritis. *Journal of the American Society of Nephrology*, 31 (6):1335–1347, June 2020. ISSN 1046-6673, 1533-3450. doi: 10.1681/ASN.2019121285.

M. Melé, P. G. Ferreira, F. Reverter, D. S. DeLuca, J. Monlong, M. Sammeth, T. R. Young, J. M. Goldmann, D. D. Pervouchine, T. J. Sullivan, R. Johnson, A. V. Segrè, S. Djebali, A. Niarchou, F. A. Wright, T. Lappalainen, M. Calvo, G. Getz, E. T. Dermitzakis, K. G. Ardlie, and R. Guigó. The human transcriptome across tissues and individuals. *Science (New York, N.Y.)*, 348(6235):660–665, May 2015. ISSN 0036-8075. doi: 10.1126/science.aaa0355.

J. Menche, A. Sharma, M. Kitsak, S. D. Ghiassian, M. Vidal, J. Loscalzo, and A.-L. Barabási. Uncovering disease-disease relationships through the incomplete interactome. *Science*, 347 (6224):1257601, 2015. Publisher: American Association for the Advancement of Science.

C. Meng, O. A. Zeleznik, G. G. Thallinger, B. Kuster, A. M. Gholami, and A. C. Culhane. Dimension reduction techniques for the integrative analysis of multi-omics data. *Briefings in Bioinformatics*, 17(4):628–641, July 2016. ISSN 1467-5463. doi: 10.1093/bib/bbv108.

M. Messina, S. Chiaretti, I. Iacobucci, S. Tavolaro, A. Lonetti, S. Santangelo, L. Elia, C. Papayannidis, F. Paoloni, A. Vitale, A. Guarini, G. Martinelli, and R. Foà. AICDA expression in BCR/ABL1-positive acute lymphoblastic leukaemia is associated with a peculiar gene expression profile. *British Journal of Haematology*, 152(6):727–732, Mar. 2011. ISSN 1365-2141. doi: 10.1111/j.1365-2141.2010.08449.x.

C. B. Messner, V. Demichev, D. Wendisch, L. Michalick, M. White, A. Freiwald, K. Textoris-Taube, S. I. Vernardis, A.-S. Egger, M. Kreidl, D. Ludwig, C. Kilian, F. Agostini, A. Zelezniak, C. Thibeault, M. Pfeiffer, S. Hippenstiel, A. Hocke, C. von Kalle, A. Campbell, C. Hayward, D. J. Porteous, R. E. Marioni, C. Langenberg, K. S. Lilley, W. M. Kuebler, M. Mülleder, C. Drosten, N. Suttorp, M. Witzenrath, F. Kurth, L. E. Sander, and M. Ralser. Ultra-High-Throughput Clinical Proteomics Reveals Classifiers of COVID-19 Infection. *Cell Systems*, 11 (1):11–24.e4, July 2020. ISSN 24054712. doi: 10.1016/j.cels.2020.05.012.

F. Mf. Epigenetic differences arise during the lifetime of monozygotic twins. *Proceedings of the National Academy of Sciences*, 102:10604–10609, 2005.

P. Mika, A. Bernstein, C. Welty, C. Knoblock, D. Vrandečić, P. Groth, N. Noy, K. Janowicz, and C. Goble. *The Semantic Web–ISWC 2014: 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part II*, volume 8797. Springer, 2014.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*, Sept. 2013a. arXiv: 1301.3781.

T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed Representations of Words and Phrases and their Compositionality. *arXiv:1310.4546 [cs, stat]*, Oct. 2013b. arXiv: 1310.4546.

M. Minsky and S. A. Papert. *Perceptrons: An Introduction to Computational Geometry*. The MIT Press, 1969. ISBN 978-0-262-34393-0. doi: 10.7551/mitpress/11301.001.0001.

B. Mirza, W. Wang, J. Wang, H. Choi, N. C. Chung, and P. Ping. Machine Learning and Integrative Analysis of Biomedical Big Data. *Genes*, 10(2):87, Jan. 2019. ISSN 2073-4425. doi: 10.3390/genes10020087.

N. Mishra, R. Schreiber, I. Stanton, and R. E. Tarjan. Clustering social networks. In *International Workshop on Algorithms and Models for the Web-Graph*, pages 56–67. Springer, 2007.

K. J. Mitchell. What is complex about complex disorders? *Genome biology*, 13(1):237, Jan. 2012. ISSN 1474-760X 1465-6906 1474-7596. doi: 10.1186/gb-2012-13-1-237. Place: England.

T. M. Mitchell. Machine learning, 1997.

S. K. Mohamed, A. Nounu, and V. Nováček. Drug target discovery using knowledge graph embeddings. In *Proceedings of the 34th ACM/SIGAPP symposium on applied computing*, pages 11–18, 2019.

S. K. Mohamed, A. Nounu, and V. Nováček. Biological applications of knowledge graph embedding models. *Briefings in Bioinformatics*, 22(2):1679–1693, Mar. 2021. ISSN 1477-4054. doi: 10.1093/bib/bbaa012.

C. Montaldo, F. Messina, I. Abbate, M. Antonioli, V. Bordoni, A. Aiello, F. Ciccosanti, F. Colavita, C. Farroni, S. Najafi Fard, E. Giombini, D. Goletti, G. Matusali, G. Rozera, M. Rueca, A. Sacchi, M. Piacentini, C. Agrati, G. M. Fimia, M. R. Capobianchi, F. N. Lauria, and G. Ippolito. Multi-omics approach to COVID-19: a domain-based literature review. *Journal of Translational Medicine*, 19(1):501, Dec. 2021. ISSN 1479-5876. doi: 10.1186/s12967-021-03168-8.

M. T. Montojo, M. Aganzo, and N. González. Huntington's Disease and Diabetes: Chronological Sequence of its Association. *Journal of Huntington's Disease*, 6(3):179–188, Sept. 2017. ISSN 18796397, 18796400. doi: 10.3233/JHD-170253.

A. A. Mortlock, D. M. Wilson, J. G. Kettle, F. W. Goldberg, and K. M. Foote. 5.02 - Selective Kinase Inhibitors in Cancer. In S. Chackalamannil, D. Rotella, and S. E. Ward, editors, *Comprehensive Medicinal Chemistry III*, pages 39–75. Elsevier, Oxford, 2017. ISBN 978-0-12-803201-5. doi: https://doi.org/10.1016/B978-0-12-409547-2.12391-1.

K. P. Murphy. *Probabilistic machine learning: an introduction*. Adaptive computation and machine learning series. The MIT Press, Cambridge, Massachusetts, 2022. ISBN 978-0-262-04682-4.

A. F. A. Musawi, S. Roy, and P. Ghosh. A Review of Link Prediction Applications in Network Biology, Dec. 2023. arXiv:2312.01275 [cs, q-bio].

W. Nelson, M. Zitnik, B. Wang, J. Leskovec, A. Goldenberg, and R. Sharan. To Embed or Not: Network Embedding as a Paradigm in Computational Biology. *Frontiers in Genetics*, 10:381, May 2019. ISSN 1664-8021. doi: 10.3389/fgene.2019.00381.

J. Newell-Price, A. J. Clark, and P. King. DNA methylation and silencing of gene expression. *Trends in Endocrinology & Metabolism*, 11(4):142–148, 2000. Publisher: Elsevier.

M. Newman. *Networks*. Oxford university press, 2018.

M. E. Newman. The structure of scientific collaboration networks. *Proceedings of the national academy of sciences*, 98(2):404–409, 2001. Publisher: National Acad Sciences.

M. E. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004. Publisher: APS.

T. K. S. Ng, C. S. H. Ho, W. W. S. Tam, E. H. Kua, and R. C.-M. Ho. Decreased serum brain-derived neurotrophic factor (BDNF) levels in patients with Alzheimer's disease (AD): a systematic review and meta-analysis. *International journal of molecular sciences*, 20(2):257, 2019. Publisher: MDPI.

D. Q. Nguyen, K. Sirts, L. Qu, and M. Johnson. STransE: a novel embedding model of entities and relationships in knowledge bases. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 460–466, 2016a. doi: 10.18653/v1/N16-1054. arXiv:1606.08140 [cs].

P. H. Nguyen, J. Giraud, C. Staedel, L. Chambonnier, P. Dubus, E. Chevret, H. Bœuf, X. Gauthereau, B. Rousseau, M. Fevre, I. Soubeyran, G. Belleannée, S. Evrard, D. Collet, F. Mégraud, and C. Varon. All-trans retinoic acid targets gastric cancer stem cells and inhibits patient-derived gastric carcinoma tumor growth. *Oncogene*, 35(43):5619–5628, Oct. 2016b. ISSN 1476-5594. doi: 10.1038/onc.2016.87.

A. C. Nica and E. T. Dermitzakis. Expression quantitative trait loci: present and future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1620):20120362, June 2013. ISSN 0962-8436. doi: 10.1098/rstb.2012.0362.

M. Nickel, K. Murphy, V. Tresp, and E. Gabrilovich. A Review of Relational Machine Learning for Knowledge Graphs. *Proceedings of the IEEE*, 104(1):11–33, Jan. 2016. ISSN 0018-9219, 1558-2256. doi: 10.1109/JPROC.2015.2483592. arXiv:1503.00759 [cs, stat].

J. Nocedal and S. J. Wright. *Numerical optimization*. Springer series in operations research. Springer, New York, 1999. ISBN 978-0-387-98793-4.

S. Nurk, S. Koren, A. Rhie, M. Rautiainen, A. V. Bzikadze, A. Mikheenko, M. R. Vollger, N. Altemose, L. Uralsky, A. Gershman, and others. The complete sequence of a human genome. *Science*, 376(6588):44–53, 2022. Publisher: American Association for the Advancement of Science.

C. Ogris, Y. Hu, J. Arloth, and N. S. Müller. Versatile knowledge guided network inference method for prioritizing key regulatory factors in multi-omics data. *Scientific Reports*, 11(1): 6806, Mar. 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-85544-4. Number: 1 Publisher: Nature Publishing Group.

J. Olkkonen, V.-P. Kouri, E. Kuusela, M. Ainola, D. Nordström, K. K. Eklund, and J. Mandelin. DEC2 Blocks the Effect of the ARNTL2/NPAS2 Dimer on the Expression of PER3 and DBP. *Journal of Circadian Rhythms*, 15, 2017. Publisher: Ubiquity Press.

C. Otte, S. M. Gold, B. W. Penninx, C. M. Pariante, A. Etkin, M. Fava, D. C. Mohr, and A. F. Schatzberg. Major depressive disorder. *Nature Reviews Disease Primers*, 2(1):16065, Sept. 2016. ISSN 2056-676X. doi: 10.1038/nrdp.2016.65.

M. Ou, P. Cui, J. Pei, Z. Zhang, and W. Zhu. Asymmetric transitivity preserving graph embedding. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1105–1114, 2016.

R. Oughtred, C. Stark, B.-J. Breitkreutz, J. Rust, L. Boucher, C. Chang, N. Kolas, L. O'Donnell, G. Leung, R. McAdam, F. Zhang, S. Dolma, A. Willems, J. Coulombe-Huntington, A. Chatraryamontri, K. Dolinski, and M. Tyers. The BioGRID interaction database: 2019 update. *Nucleic Acids Research*, 47(Database issue):D529–D541, Jan. 2019. ISSN 0305-1048. doi: 10.1093/nar/gky1079.

R. Oughtred, J. Rust, C. Chang, B.-J. Breitkreutz, C. Stark, A. Willems, L. Boucher, G. Leung, N. Kolas, F. Zhang, and others. The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Science*, 30(1):187–200, 2021. Publisher: Wiley Online Library.

K. A. Overmyer, E. Shishkova, I. J. Miller, J. Balnis, M. N. Bernstein, T. M. Peters-Clarke, J. G. Meyer, Q. Quan, L. K. Muehlbauer, E. A. Trujillo, Y. He, A. Chopra, H. C. Chieng, A. Tiwari, M. A. Judson, B. Paulson, D. R. Brademan, Y. Zhu, L. R. Serrano, V. Linke, L. A. Drake, A. P. Adam, B. S. Schwartz, H. A. Singer, S. Swanson, D. F. Mosher, R. Stewart, J. J. Coon, and A. Jaitovich. Large-Scale Multi-omic Analysis of COVID-19 Severity. *Cell Systems*, 12 (1):23–40.e7, Jan. 2021. ISSN 2405-4712. doi: 10.1016/j.cels.2020.10.003.

J. F. Padgett and C. K. Ansell. Robust Action and the Rise of the Medici, 1400-1434. *American Journal of Sociology*, 98(6):1259–1319, 1993. ISSN 00029602, 15375390. Publisher: University of Chicago Press.

L. Page, S. Brin, R. Motwani, T. Winograd, and others. The pagerank citation ranking: Bringing order to the web. *Technical report, Stanford InfoLab*, 1999. Publisher: Citeseer.

J. Z. Pan, S. Razniewski, J.-C. Kalo, S. Singhania, J. Chen, S. Dietze, H. Jabeen, J. Omeliyanenko, W. Zhang, M. Lissandrini, R. Biswas, G. de Melo, A. Bonifati, E. Vakaj, M. Dragoni, and D. Graux. Large Language Models and Knowledge Graphs: Opportunities and Challenges, Aug. 2023. arXiv:2308.06374 [cs].

H. Paulheim. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic web*, 8(3):489–508, 2017. Publisher: IOS Press.

S. Penner-Goeke and E. B. Binder. Epigenetics and depression. *Dialogues in clinical neuroscience*, 21(4):397–405, 2019. Publisher: Taylor & Francis.

B. Perozzi, R. Al-Rfou, and S. Skiena. DeepWalk: Online Learning of Social Representations. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710, Aug. 2014. doi: 10.1145/2623330.2623732. arXiv: 1403.6652.

T. Pham, T. Tran, D. Phung, and S. Venkatesh. Column Networks for Collective Classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1), Feb. 2017. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v31i1.10851.

J. Piñero, J. M. Ramírez-Anguita, J. Saüch-Pitarch, F. Ronzano, E. Centeno, F. Sanz, and L. I. Furlong. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic acids research*, 48(D1):D845–D855, 2020. Publisher: Oxford University Press.

B. T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, Jan. 1964. ISSN 0041-5553. doi: 10.1016/0041-5553(64)90137-5.

P. Pons and M. Latapy. Computing communities in large networks using random walks. In *Computer and Information Sciences-ISCIS 2005: 20th International Symposium, Istanbul, Turkey, October 26-28, 2005. Proceedings 20*, pages 284–293. Springer, 2005.

C.-H. Pui and W. E. Evans. Acute lymphoblastic leukemia. *New England Journal of Medicine*, 339(9):605–615, 1998. Publisher: Mass Medical Soc.

J. Qiu, Y. Dong, H. Ma, J. Li, K. Wang, and J. Tang. Network Embedding as Matrix Factorization: Unifying DeepWalk, LINE, PTE, and node2vec. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 459–467, Marina Del Rey CA USA, Feb. 2018. ACM. ISBN 978-1-4503-5581-0. doi: 10.1145/3159652.3159706.

M. Qu, J. Chen, L.-P. Xhonneux, Y. Bengio, and J. Tang. RNNLogic: Learning Logic Rules for Reasoning on Knowledge Graphs, July 2021. arXiv:2010.04029 [cs].

M. A. Quail, I. Kozarewa, F. Smith, A. Scally, P. J. Stephens, R. Durbin, H. Swerdlow, and D. J. Turner. A large genome center's improvements to the Illumina sequencing system. *Nature Methods*, 5(12):1005–1010, Dec. 2008. ISSN 1548-7105. doi: 10.1038/nmeth.1270.

U. N. Raghavan, R. Albert, and S. Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Physical review E*, 76(3):036106, 2007. Publisher: APS.

T. Rankinen, G. Argyropoulos, T. Rice, D. Rao, and C. Bouchard. CREB1 Is a Strong Genetic Predictor of the Variation in Exercise Heart Rate Response to Regular Exercise: The HERITAGE Family Study. *Circulation: Cardiovascular Genetics*, 3(3):294–299, June 2010. ISSN 1942-325X, 1942-3268. doi: 10.1161/CIRCGENETICS.109.925644.

E. Ravasz and A.-L. Barabási. Hierarchical organization in complex networks. *Physical review E*, 67(2):026112, 2003. Publisher: APS.

E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási. Hierarchical Organization of Modularity in Metabolic Networks. *Science*, 297(5586):1551–1555, Aug. 2002. doi: 10.1126/science.1073374. Publisher: American Association for the Advancement of Science.

D. E. Reich, M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti, D. J. Richter, T. Lavery, R. Kouyoumjian, S. F. Farhadian, R. Ward, and others. Linkage disequilibrium in the human genome. *Nature*, 411(6834):199–204, 2001. Publisher: Nature Publishing Group UK London.

B. Relja, K. Mörs, and I. Marzi. Danger signals in trauma. *Eur. J. Trauma Emerg. Surg*, 44:301, 2018.

S. Ren, Y. Shao, X. Zhao, C. S. Hong, F. Wang, X. Lu, J. Li, G. Ye, M. Yan, Z. Zhuang, C. Xu, G. Xu, and Y. Sun. Integration of Metabolomics and Transcriptomics Reveals Major Metabolic Pathways and Potential Biomarker Involved in Prostate Cancer. *Molecular & cellular proteomics : MCP*, 15(1):154–163, Jan. 2016. ISSN 1535-9484 1535-9476. doi: 10.1074/mcp.M115.052381. Place: United States.

R. Roberson-Nay, A. R. Wolen, D. M. Lapato, E. E. Lancaster, B. T. Webb, B. Verhulst, J. M. Hettema, and T. P. York. Twin Study of Early-Onset Major Depression Finds DNA Methylation Enrichment for Neurodevelopmental Genes. preprint, Genetics, Sept. 2018.

I. Rodchenkov, O. Babur, A. Luna, B. A. Aksoy, J. V. Wong, D. Fong, M. Franz, M. C. Siper, M. Cheung, M. Wrana, and others. Pathway Commons 2019 Update: integration, analysis and exploration of pathway data. *Nucleic acids research*, 48(D1):D489–D497, 2020. Publisher: Oxford University Press.

T. Rolland, M. Taşan, B. Charloteaux, S. J. Pevzner, Q. Zhong, N. Sahni, S. Yi, I. Lemmens, C. Fontanillo, R. Mosca, A. Kamburov, S. D. Ghiassian, X. Yang, L. Ghamsari, D. Balcha, B. E. Begg, P. Braun, M. Brehme, M. P. Broly, A.-R. Carvunis, D. Convery-Zupan, R. Corominas, J. Coulombe-Huntington, E. Dann, M. Dreze, A. Dricot, C. Fan, E. Franzosa, F. Gebreab, B. J. Gutierrez, M. F. Hardy, M. Jin, S. Kang, R. Kiros, G. N. Lin, K. Luck, A. MacWilliams, J. Menche, R. R. Murray, A. Palagi, M. M. Poulin, X. Rambout, J. Rasla, P. Reichert, V. Romero, E. Ruyssinck, J. M. Sahalie, A. Scholz, A. A. Shah, A. Sharma, Y. Shen, K. Spirohn, S. Tam, A. O. Tejeda, S. A. Wanamaker, J.-C. Twizere, K. Vega, J. Walsh, M. E. Cusick, Y. Xia, A.-L. Barabási, L. M. Iakoucheva, P. Aloy, J. De Las Rivas, J. Tavernier, M. A. Calderwood, D. E. Hill, T. Hao, F. P. Roth, and M. Vidal. A proteome-scale map of the human interactome network. *Cell*, 159(5):1212–1226, Nov. 2014. ISSN 0092-8674. doi: 10.1016/j.cell.2014.10.050.

O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Lecture Notes in Computer Science, pages 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4. doi: 10.1007/978-3-319-24574-4_28.

F. Rosenblatt. *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Cornell Aeronautical Laboratory. Report no. VG-1196-G-8. Spartan Books, 1962.

M. T. Rosing. ^{13}C-Depleted Carbon Microparticles in >3700-Ma Sea-Floor Sedimentary Rocks from West Greenland. *Science*, 283(5402):674–676, 1999. doi: 10.1126/science.283.5402.674. _eprint: https://www.science.org/doi/pdf/10.1126/science.283.5402.674.

M. Rosvall and C. T. Bergstrom. Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. *PloS one*, 6(4):e18209, 2011. Publisher: Public Library of Science San Francisco, USA.

C. Ruiz, M. Zitnik, and J. Leskovec. Identification of disease treatment mechanisms through the multiscale interactome. *Nature Communications*, 12(1):1796, Mar. 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-21770-8. Number: 1 Publisher: Nature Publishing Group.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning Internal Representations by Error Propagation, Parallel Distributed Processing, Explorations in the Microstructure of Cognition, ed. DE Rumelhart and J. McClelland. Vol. 1. 1986. *Biometrika*, 71:599–607, 1986.

C. D. Russell, N. I. Lone, and J. K. Baillie. Comorbidities, multimorbidity and COVID-19. *Nature medicine*, 29(2):334–343, 2023. Publisher: Nature Publishing Group US New York.

J. Y. Ryu, H. U. Kim, and S. Y. Lee. Human genes with a greater number of transcript variants tend to show biological features of housekeeping and essential genes. *Molecular BioSystems*, 11(10):2798–2807, 2015. Publisher: Royal Society of Chemistry.

N. Safari-Alighiarloo, M. Taghizadeh, M. Rezaei-Tavirani, B. Goliaei, and A. A. Peyvandi. Protein-protein interaction networks (PPI) and complex diseases. *Gastroenterology and Hepatology From Bed to Bench*, 7(1):17–31, 2014. ISSN 2008-2258.

A. Saha, Y. Kim, A. D. H. Gewirtz, B. Jo, C. Gao, I. C. McDowell, GTEx Consortium, B. E. Engelhardt, and A. Battle. Co-expression networks reveal the tissue-specific regulation of transcription and splicing. *Genome Research*, 27(11):1843–1858, Nov. 2017. ISSN 1549-5469. doi: 10.1101/gr.216721.116.

F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences*, 74(12):5463–5467, 1977. Publisher: National Acad Sciences.

T. M. Santiago-Rodriguez and E. B. Hollister. Multi 'omic data integration: A review of concepts, considerations, and approaches. *Seminars in Perinatology*, 45(6):151456, Oct. 2021. ISSN 0146-0005. doi: 10.1016/j.semperi.2021.151456.

E. E. Schadt. Molecular networks as sensors and drivers of common human diseases. *Nature*, 461(7261):218–223, Sept. 2009. ISSN 1476-4687 0028-0836. doi: 10.1038/nature08454. Place: England.

M. Schlichtkrull, T. N. Kipf, P. Bloem, R. v. d. Berg, I. Titov, and M. Welling. Modeling Relational Data with Graph Convolutional Networks, Oct. 2017. arXiv:1703.06103 [cs, stat].

W. M. Schneider, J. M. Luna, H.-H. Hoffmann, F. J. Sánchez-Rivera, A. A. Leal, A. W. Ashbrook, J. Le Pen, I. Ricardo-Lax, E. Michailidis, A. Peace, A. F. Stenzel, S. W. Lowe, M. R. MacDonald, C. M. Rice, and J. T. Poirier. Genome-Scale Identification of SARS-CoV-2 and Pan-coronavirus Host Factor Networks. *Cell*, 184(1):120–132.e14, Jan. 2021. ISSN 00928674. doi: 10.1016/j.cell.2020.12.006.

L. M. Schriml, E. Mitraka, J. Munro, B. Tauber, M. Schor, L. Nickle, V. Felix, L. Jeng, C. Bearer, R. Lichenstein, and others. Human Disease Ontology 2018 update: classification, content and workflow expansion. *Nucleic acids research*, 47(D1):D955–D962, 2019. Publisher: Oxford University Press.

R. Schulte-Sasse, S. Budach, D. Hnisz, and A. Marsico. Integration of multiomics data with graph convolutional networks to identify new cancer genes and their associated molecular mechanisms. *Nature Machine Intelligence*, 3(6):513–526, June 2021. ISSN 2522-5839. doi: 10.1038/s42256-021-00325-y. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 6 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Cancer genomics;Cellular signalling networks;Data integration;Machine learning;Tumour biomarkers Subject_term_id: cancer-genomics;cellular-signalling-networks;data-integration;machine-learning;tumour-biomarkers.

B. Schwikowski, P. Uetz, and S. Fields. A network of protein–protein interactions in yeast. *Nature Biotechnology*, 18(12):1257–1261, Dec. 2000. ISSN 1546-1696. doi: 10.1038/82360.

J. Scott. *Network analysis: A handbook*. Sage Publications, 1992.

S. Sehgal, S. Mujtaba, D. Gupta, R. Aggarwal, and R. K. Marwaha. High incidence of Epstein Barr virus infection in childhood acute lymphocytic leukemia: a preliminary study.

Indian Journal of Pathology & Microbiology, 53(1):63–67, 2010. ISSN 0974-5130. doi: 10.4103/0377-4929.59186.

N. Serin, G. H. Dihazi, A. Tayyeb, C. Lenz, G. A. Müller, M. Zeisberg, and H. Dihazi. Calreticulin Deficiency Disturbs Ribosome Biogenesis and Results in Retardation in Embryonic Kidney Development. *International Journal of Molecular Sciences*, 22(11):5858, May 2021. ISSN 1422-0067. doi: 10.3390/ijms22115858.

A. A. Shabalin. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, 28(10):1353–1358, May 2012. ISSN 1367-4811, 1367-4803. doi: 10.1093/bioinformatics/bts163.

A. Shah, B. M. John, and V. Sondhi. Acute lymphoblastic leukemia with treatment–naive Fanconi anemia. *Indian Pediatrics*, 50(5):508–510, May 2013. ISSN 0974-7559.

P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, 13(11):2498–2504, Nov. 2003. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.1239303. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

Q. Shao, Y. Wu, J. Ji, T. Xu, Q. Yu, C. Ma, X. Liao, F. Cheng, and X. Wang. Interaction Mechanisms Between Major Depressive Disorder and Non-alcoholic Fatty Liver Disease. *Frontiers in Psychiatry*, 12:711835, Dec. 2021. ISSN 1664-0640. doi: 10.3389/fpsyt.2021.711835.

C. C. Shaun Purcell. PLINK 2.0, Dec. 2019.

B. Shen, X. Yi, Y. Sun, X. Bi, J. Du, C. Zhang, S. Quan, F. Zhang, R. Sun, L. Qian, W. Ge, W. Liu, S. Liang, H. Chen, Y. Zhang, J. Li, J. Xu, Z. He, B. Chen, J. Wang, H. Yan, Y. Zheng, D. Wang, J. Zhu, Z. Kong, Z. Kang, X. Liang, X. Ding, G. Ruan, N. Xiang, X. Cai, H. Gao, L. Li, S. Li, Q. Xiao, T. Lu, Y. Zhu, H. Liu, H. Chen, and T. Guo. Proteomic and Metabolomic Characterization of COVID-19 Patient Sera. *Cell*, 182(1):59–72.e15, July 2020. ISSN 0092-8674. doi: 10.1016/j.cell.2020.05.032.

Y. Shen, J. Chen, P.-S. Huang, Y. Guo, and J. Gao. M-Walk: Learning to Walk over Graphs using Monte Carlo Tree Search, Dec. 2018. arXiv:1802.04394 [cs].

A. Shimbel. Structure in communication nets. In *Proceedings of the symposium on information networks*, pages 119–203. Polytechnic Institute of Brooklyn, 1954.

C. S. Siew, D. U. Wulff, N. M. Beckage, and Y. N. Kenett. Cognitive network science: A review of research on cognition through the lens of network representations, processes, and dynamics. *Complexity*, 2019(1):2108423, 2019. Publisher: Wiley Online Library.

M. Simeoni, T. Cavinato, D. Rodriguez, and D. Gatfield. I(nsp1)ecting SARS-CoV-2–ribosome interactions. *Communications Biology*, 4(1):1–5, June 2021. ISSN 2399-3642. doi: 10.1038/s42003-021-02265-0. Number: 1 Publisher: Nature Publishing Group.

N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. A Sparse-Group Lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, Apr. 2013. ISSN 1061-8600. doi: 10.1080/10618600.2012.681250. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10618600.2012.681250.

A. Singh, C. P. Shannon, B. Gautier, F. Rohart, M. Vacher, S. J. Tebbutt, and K.-A. L. Cao. DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics*, 35(17):3055, Sept. 2019. doi: 10.1093/bioinformatics/bty1054. Publisher: Oxford University Press.

B. Singh, P. Kaur, P. Patel, C. Nabati, S. Ayad, F. Shamoon, and M. Maroules. COVID–19 and arterial thrombosis: Report of 2 cases. *Radiology Case Reports*, 16(7):1603–1607, July 2021. ISSN 1930-0433. doi: 10.1016/j.radcr.2021.04.033.

A. Singhal. Introducing the Knowledge Graph: Things, Not Strings, 2012.

J. Smith, C. Theodoris, and E. H. Davidson. A gene regulatory network subcircuit drives a dynamic pattern of gene expression. *Science*, 318(5851):794–797, 2007. Publisher: American Association for the Advancement of Science.

S. Sookoian and C. J. Pirola. Genetic predisposition in nonalcoholic fatty liver disease. *Clinical and molecular hepatology*, 23(1):1, 2017. Publisher: Korean Association for the Study of the Liver.

N. Spataro, J. A. Rodríguez, A. Navarro, and E. Bosch. Properties of human disease genes and the role of genes linked to Mendelian disorders in complex disease aetiology. *Human Molecular Genetics*, 26(3):489–500, Feb. 2017. ISSN 0964-6906. doi: 10.1093/hmg/ddw405.

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15 (56):1929–1958, 2014. ISSN 1533-7928.

C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers. BioGRID: a general repository for interaction datasets. *Nucleic Acids Research*, 34(Database issue):D535–D539, Jan. 2006. ISSN 0305-1048. doi: 10.1093/nar/gkj109.

O. Stegle, L. Parts, R. Durbin, and J. Winn. A Bayesian Framework to Account for Complex Non-Genetic Factors in Gene Expression Levels Greatly Increases Power in eQTL Studies. *PLOS Computational Biology*, 6(5):e1000770, May 2010. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1000770. Publisher: Public Library of Science.

E. Stephenson, G. Reynolds, R. A. Botting, F. J. Calero-Nieto, M. D. Morgan, Z. K. Tuong, K. Bach, W. Sungnak, K. B. Worlock, M. Yoshida, N. Kumasaka, K. Kania, J. Engelbert, B. Olabi, J. S. Spegarova, N. K. Wilson, N. Mende, L. Jardine, L. C. S. Gardner, I. Goh, D. Horsfall, J. McGrath, S. Webb, M. W. Mather, R. G. H. Lindeboom, E. Dann, N. Huang, K. Polanski, E. Prigmore, F. Gothe, J. Scott, R. P. Payne, K. F. Baker, A. T. Hanrath, I. C. D. Schim van der Loeff, A. S. Barr, A. Sanchez-Gonzalez, L. Bergamaschi, F. Mescia, J. L. Barnes, E. Kilich, A. de Wilton, A. Saigal, A. Saleh, S. M. Janes, C. M. Smith, N. Gopee, C. Wilson, P. Coupland, J. M. Coxhead, V. Y. Kiselev, S. van Dongen, J. Bacardit, H. W. King, A. J. Rostron, A. J. Simpson, S. Hambleton, E. Laurenti, P. A. Lyons, K. B. Meyer, M. Z. Nikolić, C. J. A. Duncan, K. G. C. Smith, S. A. Teichmann, M. R. Clatworthy, J. C. Marioni, B. Göttgens, and M. Haniffa. Single-cell multi-omics analysis of the immune response in COVID-19. *Nature Medicine*, 27(5):904–916, May 2021. ISSN 1546-170X. doi: 10.1038/s41591-021-01329-2. Number: 5 Publisher: Nature Publishing Group.

S. H. Strogatz. Exploring complex networks. *Nature*, 410(6825):268–276, Mar. 2001. ISSN 0028-0836, 1476-4687. doi: 10.1038/35065725.

B. Styr and I. Slutsky. Imbalance between firing homeostasis and synaptic plasticity drives early-phase Alzheimer's disease. *Nature neuroscience*, 21(4):463–473, 2018. Publisher: Nature Publishing Group US New York.

C. Su, J. Tong, Y. Zhu, P. Cui, and F. Wang. Network embedding in biomedical data science. *Briefings in Bioinformatics*, 21(1):182–197, Jan. 2020. ISSN 1467-5463, 1477-4054. doi: 10.1093/bib/bby117.

I. Subramanian, S. Verma, S. Kumar, A. Jere, and K. Anamika. Multi-omics Data Integration, Interpretation, and Its Application. *Bioinformatics and Biology Insights*, 14:1177932219899051, Jan. 2020. ISSN 1177-9322. doi: 10.1177/1177932219899051.

J. H. Sul, L. S. Martin, and E. Eskin. Population structure in genetic studies: Confounding factors and mixed models. *PLoS genetics*, 14(12):e1007309, 2018. Publisher: Public Library of Science San Francisco, CA USA.

Y. Sun, B. Chain, S. Kaski, and J. Shawe-Taylor. Correlated Feature Selection with Extended Exclusive Group Lasso, Feb. 2020. arXiv:2002.12460 [cs, stat].

Z. Sun, Z.-H. Deng, J.-Y. Nie, and J. Tang. RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space. *arXiv:1902.10197 [cs, stat]*, Feb. 2019. arXiv: 1902.10197.

T. J. Sørensen. *A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons*. I kommission hos E. Munksgaard, 1948.

S. Tafseer, R. Gupta, R. Ahmad, S. Jain, M. Bhatia, and L. K. Gupta. Bupropion monotherapy alters neurotrophic and inflammatory markers in patients of major depressive disorder. *Pharmacology Biochemistry and Behavior*, 200:173073, 2021. Publisher: Elsevier.

D. Takai and P. A. Jones. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proceedings of the national academy of sciences*, 99(6):3740–3745, 2002. Publisher: National Acad Sciences.

X. Tan, Y. Wang, Y. Han, W. Chang, T. Su, J. Hou, D. Xu, Y. Yu, W. Ma, T. C. Thompson, and G. Cao. Genetic variation in the GSTM3 promoter confer risk and prognosis of renal cell carcinoma by reducing gene expression. *British Journal of Cancer*, 109(12):3105–3115, Dec. 2013. ISSN 0007-0920. doi: 10.1038/bjc.2013.669.

Y. Tanaka, M. Kawazu, T. Yasuda, M. Tamura, F. Hayakawa, S. Kojima, T. Ueno, H. Kiyoi, T. Naoe, and H. Mano. Transcriptional activities of DUX4 fusions in B-cell acute lymphoblastic leukemia. *Haematologica*, 103(11):e522–e526, Nov. 2018. ISSN 0390-6078. doi: 10.3324/haematol.2017.183152.

J. Tang, M. Qu, and Q. Mei. PTE: Predictive Text Embedding through Large-scale Heterogeneous Text Networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1165–1174, Aug. 2015a. doi: 10.1145/2783258.2783307. arXiv:1508.00200 [cs].

J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei. LINE: Large-scale Information Network Embedding. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1067–1077, May 2015b. doi: 10.1145/2736277.2741093. arXiv:1503.03578 [cs].

H. Tani, N. Imamachi, K. A. Salam, R. Mizutani, K. Ijiri, T. Irie, T. Yada, Y. Suzuki, and N. Akimitsu. Identification of hundreds of novel UPF1 target transcripts by direct determination of whole transcriptome stability. *RNA Biology*, 9(11):1370–1379, Nov. 2012. ISSN 1547-6286. doi: 10.4161/rna.22360.

I. Tavassoly, J. Goldfarb, and R. Iyengar. Systems biology primer: the basic methods and approaches. *Essays in Biochemistry*, 62(4):487–500, Oct. 2018. ISSN 0071-1365. doi: 10.1042/EBC20180003. _eprint: https://portlandpress.com/essaysbiochem/article-pdf/62/4/487/887339/ebc-2018-0003c.pdf.

M. Z. Tay, C. M. Poh, L. Rénia, P. A. MacAry, and L. F. P. Ng. The trinity of COVID-19: immunity, inflammation and intervention. *Nature Reviews. Immunology*, pages 1–12, Apr. 2020. ISSN 1474-1733. doi: 10.1038/s41577-020-0311-8.

A. C. Tecalco-Cruz, J. O. Ramírez-Jarquín, M. E. Alvarez-Sánchez, and J. Zepeda-Cervantes. Epigenetic basis of Alzheimer disease. *World Journal of Biological Chemistry*, 11(2):62–75, Sept. 2020. ISSN 1949-8454. doi: 10.4331/wjbc.v11.i2.62.

A. Terracciano. Genome-wide association scan of trait depression. *Biol. Psychiat*, 68:811–817, 2010.

T. Terwilliger and M. Abdul-Hay. Acute lymphoblastic leukemia: a comprehensive review and 2017 update. *Blood Cancer Journal*, 7(6):e577–e577, June 2017. ISSN 2044-5385. doi: 10.1038/bcj.2017.53. Number: 6 Publisher: Nature Publishing Group.

The COVID-19 Host Genetics Initiative. The COVID-19 Host Genetics Initiative, a global initiative to elucidate the role of host genetic factors in susceptibility and severity of the SARS-CoV-2 virus pandemic. *European Journal of Human Genetics*, 28(6):715–718, June 2020. ISSN 1018-4813, 1476-5438. doi: 10.1038/s41431-020-0636-6.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996. Publisher: Wiley Online Library.

T. Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the Gradient by a Running Average of Its Recent Magnitude. *COURSERA: Neural Networks for Machine Learning, 4, 26-31*, 2012.

R. Tiwari, A. R. Mishra, A. Gupta, and D. Nayak. Structural similarity-based prediction of host factors associated with SARS-CoV-2 infection and pathogenesis. *Journal of Biomolecular Structure and Dynamics*, pages 1–12, Jan. 2021. ISSN 0739-1102, 1538-0254. doi: 10.1080/07391102.2021.1874532.

T. Trouillon, J. Welbl, S. Riedel, E. Gaussier, and G. Bouchard. Complex Embeddings for Simple Link Prediction, June 2016. arXiv:1606.06357 [cs, stat].

M. Uhlen, L. Fagerberg, B. M. Hallstroem, C. Lindskog, P. Oksvold, A. Mardinoglu, A. Sivertsson, C. Kampf, E. Sjoestedt, A. Asplund, I. Olsson, K. Edlund, E. Lundberg, S. Navani, C. A.-K. Szigyarto, J. Odeberg, D. Djureinovic, J. O. Takanen, S. Hober, T. Alm, P.-H. Edqvist, H. Berling, H. Tegel, J. Mulder, J. Rockberg, P. Nilsson, J. M. Schwenk, M. Hamsten, K. von Feilitzen, M. Forsberg, L. Persson, F. Johansson, M. Zwahlen, G. von Heijne, J. Nielsen, and F. Ponten. Tissue-based map of the human proteome. *Science*, 347(6220):1260419, Jan. 2015. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1260419.

K. Urbanek, D. Torella, F. Sheikh, A. De Angelis, D. Nurzynska, F. Silvestri, C. A. Beltrami, R. Bussani, A. P. Beltrami, F. Quaini, R. Bolli, A. Leri, J. Kajstura, and P. Anversa. Myocardial regeneration by activation of multipotent cardiac stem cells in ischemic heart failure. *Proceedings of the National Academy of Sciences*, 102(24):8692–8697, June 2005. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0500169102.

L. A. Urry, M. L. Cain, S. A. Wasserman, P. V. Minorsky, and J. B. Reece. *Campbell Biologie*. Pearson Deutschland, 2019. ISBN 978-3-86894-366-5.

M. A. Valencia-Sanchez, J. Liu, G. J. Hannon, and R. Parker. Control of translation and mRNA degradation by miRNAs and siRNAs. *Genes & development*, 20(5):515–524, 2006. Publisher: Cold Spring Harbor Lab.

C. H. Van Dyck, C. J. Swanson, P. Aisen, R. J. Bateman, C. Chen, M. Gee, M. Kanekiyo, D. Li, L. Reyderman, S. Cohen, and others. Lecanemab in early Alzheimer's disease. *New England Journal of Medicine*, 388(1):9–21, 2023. Publisher: Mass Medical Soc.

C. Varallo-Rodriguez, C. W. Freyer, E. P. Ontiveros, E. A. Griffiths, E. S. Wang, and M. Wetzler. Bosutinib for the Treatment of Philadelphia Chromosome-Positive Leukemias. *Expert opinion on orphan drugs*, 3(5):599–608, 2015. ISSN 2167-8707. doi: 10.1517/21678707.2015.1036027.

K. Z. Vardakas, I. I. Siempos, and M. E. Falagas. Diabetes mellitus as a risk factor for nosocomial pneumonia and associated mortality. *Diabetic Medicine*, 24(10): 1168–1171, 2007. ISSN 1464-5491. doi: 10.1111/j.1464-5491.2007.02234.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1464-5491.2007.02234.x.

P. Varinthra and I. Liu. Molecular basis for the association between depression and circadian rhythm. *Ci Ji Yi Xue Za Zhi*, 31:67–72, 2019.

S. Vashishth, S. Sanyal, V. Nitin, and P. Talukdar. Composition-based Multi-Relational Graph Convolutional Networks, Jan. 2020. arXiv:1911.03082 [cs, stat].

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

E. Vegeto, A. Villa, S. Della Torre, V. Crippa, P. Rusmini, R. Cristofani, M. Galbiati, A. Maggi, and A. Poletti. The role of sex and sex hormones in neurodegenerative diseases. *Endocrine reviews*, 41(2):273–319, 2020. Publisher: Oxford University Press US.

P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio. Graph Attention Networks, Feb. 2018. arXiv:1710.10903 [cs, stat].

U. Võsa, A. Claringbould, H.-J. Westra, M. J. Bonder, P. Deelen, B. Zeng, H. Kirsten, A. Saha, R. Kreuzhuber, S. Yazar, H. Brugge, R. Oelen, D. H. de Vries, M. G. P. van der Wijst, S. Kasela, N. Pervjakova, I. Alves, M.-J. Favé, M. Agbessi, M. W. Christiansen, R. Jansen, I. Seppälä, L. Tong, A. Teumer, K. Schramm, G. Hemani, J. Verlouw, H. Yaghootkar, R. Sönmez Flitman, A. Brown, V. Kukushkina, A. Kalnapenkis, S. Rüeger, E. Porcu, J. Kronberg, J. Kettunen, B. Lee, F. Zhang, T. Qi, J. A. Hernandez, W. Arindrarto, F. Beutner, J. Dmitrieva, M. Elansary, B. P. Fairfax, M. Georges, B. T. Heijmans, A. W. Hewitt, M. Kähönen, Y. Kim, J. C. Knight, P. Kovacs, K. Krohn, S. Li, M. Loeffler, U. M. Marigorta, H. Mei, Y. Momozawa, M. Müller-Nurasyid, M. Nauck, M. G. Nivard, B. W. J. H. Penninx, J. K. Pritchard,

O. T. Raitakari, O. Rotzschke, E. P. Slagboom, C. D. A. Stehouwer, M. Stumvoll, P. Sullivan, P. A. C. 't Hoen, J. Thiery, A. Tönjes, J. van Dongen, M. van Iterson, J. H. Veldink, U. Völker, R. Warmerdam, C. Wijmenga, M. Swertz, A. Andiappan, G. W. Montgomery, S. Ripatti, M. Perola, Z. Kutalik, E. Dermitzakis, S. Bergmann, T. Frayling, J. van Meurs, H. Prokisch, H. Ahsan, B. L. Pierce, T. Lehtimäki, D. I. Boomsma, B. M. Psaty, S. A. Gharib, P. Awadalla, L. Milani, W. H. Ouwehand, K. Downes, O. Stegle, A. Battle, P. M. Visscher, J. Yang, M. Scholz, J. Powell, G. Gibson, T. Esko, and L. Franke. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nature Genetics*, 53(9):1300–1310, Sept. 2021. ISSN 1546-1718. doi: 10.1038/s41588-021-00913-z. Number: 9 Publisher: Nature Publishing Group.

A. J. M. Walhout and M. Vidal. High-Throughput Yeast Two-Hybrid Assays for Large-Scale Protein Interaction Mapping. *Methods*, 24(3):297–306, 2001. ISSN 1046-2023. doi: https://doi.org/10.1006/meth.2001.1190.

Y.-W. Wan, R. Al-Ouran, C. G. Mangleburg, T. M. Perumal, T. V. Lee, K. Allison, V. Swarup, C. C. Funk, C. Gaiteri, M. Allen, and others. Meta-analysis of the Alzheimer's disease human brain transcriptome and functional dissection in mouse models. *Cell reports*, 32(2), 2020. Publisher: Elsevier.

C.-L. Wang. VPS35 regulates developing mouse hippocampal neuronal morphogenesis by promoting retrograde trafficking of BACE1. *Biol. Open*, 1:1248–1257, 2012.

Q. Wang, M. Li, X. Wang, N. Parulian, G. Han, J. Ma, J. Tu, Y. Lin, R. H. Zhang, W. Liu, A. Chauhan, Y. Guan, B. Li, R. Li, X. Song, Y. Fung, H. Ji, J. Han, S.-F. Chang, J. Pustejovsky, J. Rah, D. Liem, A. ELsayed, M. Palmer, C. Voss, C. Schneider, and B. Onyshkevych. COVID-19 Literature Knowledge Graph Construction and Drug Repurposing Report Generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 66–77, Online, 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-demos.8.

W. Wang, L. Min, X. Qiu, X. Wu, C. Liu, J. Ma, D. Zhang, and L. Zhu. Biological function of long non-coding RNA (LncRNA) Xist. *Frontiers in cell and developmental biology*, 9:645647, 2021b. Publisher: Frontiers Media SA.

W.-Y. Wang, M.-S. Tan, J.-T. Yu, and L. Tan. Role of pro-inflammatory cytokines released from microglia in Alzheimer's disease. *Annals of Translational Medicine*, 3(10):136, June 2015. ISSN 2305-5839. doi: 10.3978/j.issn.2305-5839.2015.03.49.

X. Wang and G. Sukthankar. Link Prediction in Heterogeneous Collaboration Networks. In R. Missaoui and I. Sarr, editors, *Social Network Analysis - Community Detection and Evolution*, Lecture Notes in Social Networks, pages 165–192. Springer International Publishing, Cham, 2014. ISBN 978-3-319-12188-8. doi: 10.1007/978-3-319-12188-8_8.

Y. Wang, Z. You, L. Li, and Z. Chen. A survey of current trends in computational predictions of protein-protein interactions. *Frontiers of Computer Science*, 14:1–12, 2020. Publisher: Springer.

Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57–63, 2009. Publisher: Nature Publishing Group UK London.

Z. Wang, J. Zhang, J. Feng, and Z. Chen. Knowledge Graph Embedding by Translating on Hyperplanes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 28(1), June 2014. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v28i1.8870.

Z. Wang, Z. Wang, B. Srinivasan, V. N. Ioannidis, H. Rangwala, and R. Anubhai. BioBridge: Bridging Biomedical Foundation Models via Knowledge Graphs, Jan. 2024. arXiv:2310.03320 [cs].

M. D. Ward, K. Stovel, and A. Sacks. Network analysis and political science. *Annual Review of Political Science*, 14:245–264, 2011. Publisher: Annual Reviews.

S. Wasserman and K. Faust. Social network analysis: Methods and applications. *Cambridge university press*, 1994. Publisher: Cambridge university press.

D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684): 440–442, 1998. Publisher: Nature Publishing Group.

J. Wei, M. M. Alfajaro, P. C. DeWeirdt, R. E. Hanna, W. J. Lu-Culligan, W. L. Cai, M. S. Strine, S.-M. Zhang, V. R. Graziano, C. O. Schmitz, J. S. Chen, M. C. Mankowski, R. B. Filler, N. G. Ravindra, V. Gasque, F. J. de Miguel, A. Patil, H. Chen, K. Y. Oguntuyo, L. Abriola, Y. V. Surovtseva, R. C. Orchard, B. Lee, B. D. Lindenbach, K. Politi, D. van Dijk, C. Kadoch, M. D. Simon, Q. Yan, J. G. Doench, and C. B. Wilen. Genome-wide CRISPR Screens Reveal Host Factors Critical for SARS-CoV-2 Infection. *Cell*, 184(1):76–91.e13, Jan. 2021. ISSN 00928674. doi: 10.1016/j.cell.2020.10.028.

S. S. Weinreich, R. Mangon, J. Sikkens, M. E. Teeuw, and M. Cornel. Orphanet: a European database for rare diseases. *Nederlands tijdschrift voor geneeskunde*, 152(9):518–519, 2008.

J. Weinstein. The cancer genome atlas pan-cancer analysis project. *Nat. Genet*, 45:1113–1120, 2013.

R. J. Williams and N. D. Martinez. Simple rules yield complex food webs. *Nature*, 404(6774): 180–183, 2000. Publisher: Nature Publishing Group UK London.

D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, and others. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic acids research*, 46(D1):D1074–D1082, 2018. Publisher: Oxford University Press.

P. Wu, D. Chen, W. Ding, P. Wu, H. Hou, Y. Bai, Y. Zhou, K. Li, S. Xiang, P. Liu, J. Ju, E. Guo, J. Liu, B. Yang, J. Fan, L. He, Z. Sun, L. Feng, J. Wang, T. Wu, H. Wang, J. Cheng, H. Xing, Y. Meng, Y. Li, Y. Zhang, H. Luo, G. Xie, X. Lan, Y. Tao, J. Li, H. Yuan, K. Huang, W. Sun, X. Qian, Z. Li, M. Huang, P. Ding, H. Wang, J. Qiu, F. Wang, S. Wang, J. Zhu, X. Ding, C. Chai, L. Liang, X. Wang, L. Luo, Y. Sun, Y. Yang, Z. Zhuang, T. Li, L. Tian, S. Zhang, L. Zhu, A. Chang, L. Chen, Y. Wu, X. Ma, F. Chen, Y. Ren, X. Xu, S. Liu, J. Wang, H. Yang, L. Wang, C. Sun, D. Ma, X. Jin, and G. Chen. The trans-omics landscape of COVID-19. *Nature Communications*, 12(1):4543, July 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-24482-1. Number: 1 Publisher: Nature Publishing Group.

Y.-L. Wu, Z.-J. Lin, C.-C. Li, X. Lin, S.-K. Shan, B. Guo, M.-H. Zheng, F. Li, L.-Q. Yuan, and Z.-h. Li. Epigenetic regulation in metabolic diseases: mechanisms and advances in clinical study. *Signal Transduction and Targeted Therapy*, 8(1):1–27, Mar. 2023. ISSN 2059-3635. doi: 10.1038/s41392-023-01333-7. Publisher: Nature Publishing Group.

M. A. Wörheide, J. Krumsiek, S. Nataf, K. Nho, A. K. Greenwood, T. Wu, K. Huynh, P. Weinisch, W. Römisch-Margl, N. Lehner, and others. An integrated molecular atlas of Alzheimer's disease. *medRxiv*, pages 2021–09, 2021. Publisher: Cold Spring Harbor Laboratory Press.

Y. Xiao, Y. Hou, H. Zhou, G. Diallo, M. Fiszman, J. Wolfson, L. Zhou, H. Kilicoglu, Y. Chen, C. Su, H. Xu, W. G. Mantyh, and R. Zhang. Repurposing non-pharmacological interventions for Alzheimer's disease through link prediction on biomedical literature. *Scientific Reports*, 14 (1):8693, Apr. 2024. ISSN 2045-2322. doi: 10.1038/s41598-024-58604-8. Publisher: Nature Publishing Group.

W. Xiong, T. Hoang, and W. Y. Wang. DeepPath: A Reinforcement Learning Method for Knowledge Graph Reasoning. In M. Palmer, R. Hwa, and S. Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 564–573, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1060.

K. Xu, C. Li, Y. Tian, T. Sonobe, K.-i. Kawarabayashi, and S. Jegelka. Representation Learning on Graphs with Jumping Knowledge Networks, June 2018. arXiv:1806.03536 [cs, stat].

K. Xu, W. Hu, J. Leskovec, and S. Jegelka. How Powerful are Graph Neural Networks?, Feb. 2019. arXiv:1810.00826 [cs, stat].

J. Yan, S. L. Risacher, L. Shen, and A. J. Saykin. Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data. *Briefings in Bioinformatics*, 19 (6):1370–1381, June 2017. ISSN 1467-5463. doi: 10.1093/bib/bbx066.

B. Yang, W.-t. Yih, X. He, J. Gao, and L. Deng. Embedding Entities and Relations for Learning and Inference in Knowledge Bases, Aug. 2015. arXiv:1412.6575 [cs].

H. Yang, Z. Lin, and M. Zhang. Rethinking Knowledge Graph Evaluation Under the Open-World Assumption, Sept. 2022. arXiv:2209.08858 [cs, stat].

J. You, Z. Ying, and J. Leskovec. Design space for graph neural networks. *Advances in Neural Information Processing Systems*, 33:17009–17021, 2020.

M. Yuan and Y. Lin. Model Selection and Estimation in Regression with Grouped Variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(1):49–67, Feb. 2006. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2005.00532.x.

M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. R. Salakhutdinov, and A. J. Smola. Deep Sets. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

M. F. Zamakhchari, C. Sima, K. Sama, N. Fine, M. Glogauer, T. E. Van Dyke, and R. Gyurko. Lack of p47phox in Akita Diabetic Mice Is Associated with Interstitial Pneumonia, Fibrosis, and Oral Inflammation. *The American Journal of Pathology*, 186(3):659–670, Mar. 2016. ISSN 0002-9440. doi: 10.1016/j.ajpath.2015.10.026.

A. S. Zannas, J. Arloth, T. Carrillo-Roa, S. Iurato, S. Röh, K. J. Ressler, C. B. Nemeroff, A. K. Smith, B. Bradley, C. Heim, A. Menke, J. F. Lange, T. Brückl, M. Ising, N. R. Wray, A. Erhardt, E. B. Binder, and D. Mehta. Lifetime stress accelerates epigenetic aging in an

urban, African American cohort: relevance of glucocorticoid signaling. *Genome Biology*, 16 (1):266, Dec. 2015. ISSN 1474-760X. doi: 10.1186/s13059-015-0828-5.

D. Zhang, J. Yin, X. Zhu, and C. Zhang. Network Representation Learning: A Survey, July 2018. arXiv:1801.05852 [cs, stat].

J. Zhang, K. McCastlain, H. Yoshihara, B. Xu, Y. Chang, M. L. Churchman, G. Wu, Y. Li, L. Wei, I. Iacobucci, Y. Liu, C. Qu, J. Wen, M. Edmonson, D. Payne-Turner, K. B. Kaufmann, S.-i. Takayanagi, E. Wienholds, E. Waanders, P. Ntziachristos, S. Bakogianni, J. Wang, I. Aifantis, K. G. Roberts, J. Ma, G. Song, J. Easton, H. L. Mulder, X. Chen, S. Newman, X. Ma, M. Rusch, P. Gupta, K. Boggs, B. Vadodaria, J. Dalton, Y. Liu, M. L. Valentine, L. Ding, C. Lu, R. S. Fulton, L. Fulton, Y. Tabib, K. Ochoa, M. Devidas, D. Pei, C. Cheng, J. Yang, W. E. Evans, M. V. Relling, C.-H. Pui, S. Jeha, R. C. Harvey, I.-M. L. Chen, C. L. Willman, G. Marcucci, C. D. Bloomfield, J. Kohlschmidt, K. Mrózek, E. Paietta, M. S. Tallman, W. Stock, M. C. Foster, J. Racevskis, J. M. Rowe, S. Luger, S. M. Kornblau, S. A. Shurtleff, S. C. Raimondi, E. R. Mardis, R. K. Wilson, J. E. Dick, S. P. Hunger, M. L. Loh, J. R. Downing, and C. G. Mullighan. DEREGULATION OF DUX4 AND ERG IN ACUTE LYMPHOBLASTIC LEUKEMIA. *Nature genetics*, 48(12):1481–1489, Dec. 2016. ISSN 1061-4036. doi: 10.1038/ng.3691.

K. Zhang, M. Wu, Y. Liu, Y. Feng, and J. Zheng. KR4SL: knowledge graph reasoning for explainable prediction of synthetic lethality. *Bioinformatics*, 39(Supplement_1):i158–i167, June 2023. ISSN 1367-4811. doi: 10.1093/bioinformatics/btad261.

S. Zhang, C.-C. Liu, W. Li, H. Shen, P. W. Laird, and X. J. Zhou. Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Research*, 40(19):9379–9391, Oct. 2012. ISSN 0305-1048. doi: 10.1093/nar/gks725.

Y. Zhang and Q. Yao. Knowledge Graph Reasoning with Relational Digraph. In *Proceedings of the ACM Web Conference 2022*, pages 912–924, Apr. 2022. doi: 10.1145/3485447.3512008. arXiv:2108.06040 [cs].

Y. Zhang, Z. Chen, L. Guo, Y. Xu, W. Zhang, and H. Chen. Making Large Language Models Perform Better in Knowledge Graph Completion, Apr. 2024. arXiv:2310.06671 [cs].

Y. Zhao, Y. Zhu, H. Wang, and C. Ji. Case Report: Successful Treatment of Cutaneous Squamous Cell Carcinoma in Three Patients With a Combination of Acitretin and Clarithromycin. *Frontiers in Oncology*, 11, 2021. ISSN 2234-943X.

X. Zheng, L. Chen, H. Xia, H. Wei, Q. Lou, M. Li, T. Li, and L. Luo. Transgenerational epimutations induced by multi-generation drought imposition mediate rice plant's adaptation to drought condition. *Scientific reports*, 7(1):39843, 2017. Publisher: Nature Publishing Group UK London.

Z. Zhu, Z. Zhang, L.-P. Xhonneux, and J. Tang. Neural Bellman-Ford Networks: A General Graph Neural Network Framework for Link Prediction. In *Advances in Neural Information Processing Systems*, volume 34, pages 29476–29490. Curran Associates, Inc., 2021.

M. Zitnik and J. Leskovec. Predicting multicellular function through multi-layer tissue networks. *Bioinformatics*, 33(14):i190–i198, July 2017. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btx252.

M. Zitnik, M. Agrawal, and J. Leskovec. Modeling Polypharmacy Side Effects with Graph Convolutional Networks. *Bioinformatics*, page 9, 2018.

H. Zou and T. Hastie. Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(2):301–320, 2005. ISSN 13697412, 14679868. Publisher: [Royal Statistical Society, Wiley].

List of Figures

List of Tables