

Drug Repurposing through Network Medicine

Sepideh Sadegh

Vollständiger Abdruck der von der TUM School of Life Sciences der Technischen Universität München zur Erlangung einer
Doktorin der Naturwissenschaften (Dr. rer. nat.)
genehmigten Dissertation.

Vorsitz: Prof. Dr. Mathias Wilhelm

Prüfende der Dissertation:

1. Prof. Dr. Jan Baumbach
2. Prof. Markus List, Ph.D.
3. Prof. Dr. Olga Kalinina

Die Dissertation wurde am 11.12.2023 bei der Technischen Universität München eingereicht und durch die TUM School of Life Sciences am 01.05.2024 angenommen.

Acknowledgements

Throughout my time as a PhD student, many people played a crucial role in discussions of the project and supporting me to keep going when everything felt not working well. I am very grateful for all of your support even if you are not named here.

First, I would like to express my gratitude to my supervisor, Prof. Jan Baumbach, for giving me the opportunity to work on the exciting topic of drug repurposing under the impactful EU project, REPO-TRIAL.

I cannot thank Prof. David Blumenthal enough for his great support. He was not officially my supervisor but he was de facto guiding me especially during the second half of my project. I would like to thank Prof. Tim Kacprowski, the initial support I received from him paved my way through the PhD journey.

I would like to forward my thankfulness to past and present members of the Exbio lab, in particular, Dr. Markus List, Martina and Léonie. For inspirational discussions and support, thank you to Gihan, Marisol, Julian, and Andi.

Working in a collaborative environment with highly-driven colleagues from different universities across Europe, taught me a lot not only on a scientific level but also on a personal level. I highly appreciate the chance I had to work with Prof. Harald Schmidt in the REPO-TRIAL project. A special thank you to Prof. Anil Wipat, James and Elisa, our collaborators from the New Castle University. Working with you, although always remotely, was a very fruitful and enriching experience for me. I would like to thank the computational biomedicine lab from the University of Southern Denmark for hosting me. This amazing experience lured me into moving to Denmark! I'm grateful to my friend Afsaneh for helping make my transition to the new country smoother.

A big thanks to my fiancé, Richard, I could not have hoped for a better partner during all the frustrating periods of a PhD. A special thank you also goes to my kind family-in-law, Margit and Walter, who supported me as if I were their own child.

At last but definitely not the least, I would like to thank my family, Fereshteh, Mohammad, and Sepehr, for believing in me and always being there for me despite the physical distance.

Abstract

In recent decades, despite substantial advancements in science and technology, drug research and development efficiency has stagnated. Consequently, drug repurposing (DR) - seeking new applications for existing drugs - has gained prominence as an alternative strategy. DR offers key benefits, notably reduced time and costs, making it attractive especially in urgent situations like the COVID-19 pandemic. Computational methods based on network medicine principles expedite DR by narrowing the search for therapeutics.

Advancements in high-throughput omics technologies have empowered researchers to comprehensively study DNA, RNA, proteins, and molecules, leading to insights into genetic risk factors and disease mechanisms. Integrating data from different molecular levels offers opportunities for personalized treatments. However, the complexity of noisy molecular profiling data poses challenges in distinguishing causal relationships, motivating the incorporation of prior information, like protein-protein interaction networks, enabling integrative analyses to discover pathomechanisms.

Network medicine, which models complex biological systems by incorporating diverse data types, helps identify disease modules—collections of molecular entities explaining disease phenotypes and pathomechanisms. These approaches can further aid DR by identifying already approved drugs targeting the discovered disease modules. However, this necessitates an integrated knowledge base consolidating scattered databases of drug- and disease-related data.

In this cumulative thesis, I present the development of two unified systems medicine platforms for DR, by combining complex network-medicine algorithms with a harmonized knowledge base.

The first publication introduces CoVex, a platform designed for COVID-19 DR. CoVex integrates virus-human interaction data, protein-protein interaction networks, and drug information, employing network medicine algorithms to predict novel drug targets and candidates. CoVex highlights the utility of expert knowledge integration, offering a user-friendly web tool to explore molecular mechanisms.

The second publication introduces NeDRex, a generalized DR platform extending CoVex. NeDRex constructs heterogeneous biological networks enriched with disease and drug data, enabling the mining of candidate disease mechanisms while engaging expert knowledge. The platform prioritizes drugs targeting proteins involved in these mechanisms, providing a versatile tool applicable to practically any disease.

The third publication offers a comprehensive review of computational approaches for COVID-19 DR, categorizing these approaches into virus-targeting and host-

targeting strategies and offering insights into relevant data resources. It emphasizes the need for a unified DR strategy to enhance preparedness for future outbreaks.

The fourth publication delves into the critical issue of inadequate disease definitions in network medicine studies, which can introduce bias and hinder progress. Through global- and local-scale similarity analyses of complex networks constructed from disease and drug association data, this study uncovers the risks associated with mechanistically inadequate disease definitions. It underscores the importance of complementing publicly available disease association data with well-characterized patient cohort data for more reliable network medicine analyses.

In sum, this dissertation presents network medicine platforms for DR, underpinned by the integration of a multitude of databases, expert knowledge, and novel computational approaches. All platforms aim for high accessibility for biomedical researchers, facilitating analyses and interpretation of predictions. This dissertation explores challenges in the field, from pandemic responses to the refinement of disease definitions, ultimately contributing to the advancement of precision medicine and therapeutic discovery.

Kurzfassung

In den letzten Jahrzehnten stagnierte die Effizienz der Arzneimittelforschung und -entwicklung trotz erheblicher Fortschritte in Wissenschaft und Technologie. Dadurch hat das so genannte „Drug Repurposing“ (DR), also die Suche nach neuen Behandlungsanwendungen für bestehende Wirkstoffe, als alternative Strategie zur Medikamententwicklung an Bedeutung gewonnen. DR bietet insbesondere durch einen geringeren Zeit- und Kostenaufwand entscheidende Vorteile und macht es gerade in dringenden Situationen wie der COVID-19-Pandemie besonders attraktiv. Aufbauend auf den Prinzipien der Netzwerkmedizin beschleunigen computergestützte Methoden das DR, da sie den Suchraum für neu Therapeutika stark einschränken.

Fortschritte in der Hochdurchsatz-Omics-Technologie ermöglichen heutzutage die umfassende Untersuchung der DNA, RNA, Proteine und Moleküle und decken dadurch genetische Risikofaktoren und Krankheitsmechanismen auf. Die Integration von Daten aus verschiedenen molekularen Ebenen ermöglicht schließlich personalisierte Behandlungen. Nichtsdestotrotz stellt das Signalrauschen in den molekularen Profilen ein erhebliches Problem bei der Unterscheidung kausaler Zusammenhänge dar und begründet die Einbeziehung von Vorwissen in Form von beispielsweise Protein-Protein-Interaktionsnetzwerke, was integrative Analysen zur Entdeckung von Pathomechanismen ermöglicht.

Die Netzwerkmedizin modelliert diese komplexen biologische Systeme und hilft dadurch bei der Identifizierung von Krankheitsmodulen – eine Sammlung von molekularen Entitäten, die zusammen die Krankheitsphänotypen und Pathomechanismen erklären. Diese wiederum können das DR erheblich unterstützen, indem bereits zugelassene Medikamente, die auf entdeckte Krankheitsmodule abzielen, identifiziert werden. Dies erfordert jedoch eine integrierte Wissensbasis, die verschiedenste Datenbanken mit Arzneimittel- und Krankheitsbezogenen Daten konsolidiert.

In dieser kumulativen Arbeit präsentiere ich die Entwicklung zweier ganzheitlicher systemmedizinischer Plattformen für das DR, die komplexe netzwerkmedizinische Algorithmen mit einer harmonisierten Wissensbasis kombinieren.

Die erste Veröffentlichung behandelt CoVex, eine Plattform die speziell für das DR für COVID-19 entwickelt wurde. CoVex integriert Daten zur Virus-Mensch-Interaktion, Protein-Protein-Interaktionsnetzwerke und Arzneimittelinformationen und verwendet Netzwerkmedizin-Algorithmen, um neue Medikamentenziele und Kandidaten vorherzusagen. CoVex unterstreicht den Nutzen der Integration von Expertenwissen und bietet ein benutzerfreundliches Web-Tool zur Erforschung molekularer Mechanismen.

Die zweite Veröffentlichung stellt mit NeDRex eine verallgemeinerte Version von CoVex für das DR vor. NeDRex verarbeitet heterogene biologische Netzwerke, angereichert mit Krankheits- und Arzneimitteldaten, und ermöglicht so die Entdeckung möglicher Krankheitsmechanismen unter Einbeziehung von Expertenwissen. Die Plattform priorisiert Medikamente, die auf an den Mechanismen beteiligten Proteine abzielen, und bietet so ein vielseitiges Werkzeug, das praktisch auf jede Krankheit anwendbar ist.

Die dritte Veröffentlichung bietet einen umfassenden Überblick über computerunterstützte Ansätze für das COVID-19 DR und kategorisiert diese Ansätze in Viren-Targeting- und Host-Targeting-Strategien und bietet Einblicke in relevante Datenquellen. Die Arbeit betont die Notwendigkeit einer einheitlichen DR-Strategie, um besser auf künftige Ausbrüche vorbereitet zu sein.

Die vierte Veröffentlichung befasst sich mit dem kritischen Problem unzureichender und ungenauer Krankheitsdefinitionen in Netzwerkmedizinstudien, die zu Verzerrungen führen und dadurch Fortschrittshemmend wirken. Durch globale und lokale Ähnlichkeitsanalysen komplexer Netzwerke, die aus Krankheits- und Arzneimittelassoziationsdaten erstellt wurden, deckt diese Studie die Risiken von mechanistisch unzureichenden Krankheitsdefinitionen auf. Die Arbeit unterstreicht die Bedeutung der Ergänzung öffentlich verfügbarer Krankheitsassoziationsdaten durch gut charakterisierte Patientenkohortendaten für zuverlässigere Netzwerkmedizin-Analysen.

Insgesamt stellt diese Dissertation neue Plattformen für die Anwendung von Netzwerkmedizin für das DR vor, die auf der Integration einer Vielzahl von Datenbanken, Expertenwissen und neuartigen Algorithmen basieren. Alle Plattformen sind besonders benutzerfreundlich für biomedizinische Forscher gestaltet, um Analysen und Interpretation von Vorhersagen zu erleichtern. Diese Dissertation bearbeitet auf vielschichtige Art und Weise die Anwendung der Netzwerkmedizin, von der schnellen Reaktion auf eine Pandemielage bis hin zur Verfeinerung von Krankheitsdefinitionen, und trägt letztendlich zur Weiterentwicklung der Präzisionsmedizin und der Therapieentdeckung bei.

Publication Record

Publications presented in this thesis (first author and published):

- 1) **Sadegh S**†, Matschinske J†, Blumenthal DB, Galindez G, Kacprowski T, List M, Nasirigerdeh R, Oubounyt M, Pichlmair A, Rose TD, Salgado-Albarrán M, Späth J, Stukalov A, Wenke NK, Yuan K, Pauling JK & Baumbach J (2020). “Exploring the SARS-CoV-2 virus-host-drug interactome for drug repurposing.” *Nature communications*, 11(1), 3518; <https://doi.org/10.1038/s41467-020-17189-2>
- 2) **Sadegh S**†, Skelton J†, Anastasi E, Bernett J, Blumenthal DB, Galindez G, Salgado-Albarrán M, Lazareva O, Flanagan K, Cockell S, Nogales C, Casas AI, Schmidt HHHW, Baumbach J, Wipat A & Kacprowski T (2021). “Network medicine for disease module identification and drug repurposing with the NeDRex platform.” *Nature communications*, 12(1), 6848; <https://doi.org/10.1038/s41467-021-27138-2>
- 3) Galindez G†, Matschinske J†, Rose TD†, **Sadegh S**†, Salgado-Albarrán M†, Späth J†, Baumbach J & Pauling JK (2021). “Lessons from the COVID-19 pandemic for advancing computational drug repurposing strategies.” *Nature Computational Science*, 1(1), 33-41; <https://doi.org/10.1038/s43588-020-00007-6>
- 4) **Sadegh S**, Skelton J, Anastasi E, Maier A, Adamowicz K, Möller A, Kriege NM, Kronberg J, Haller T, Kacprowski T, Wipat A, Baumbach J, Blumenthal DB (2023). “Lacking mechanistic disease definitions and corresponding association data hamper progress in network medicine and beyond.” *Nature Communications*, 14(1), 1662; <https://doi.org/10.1038/s41467-023-37349-4>

† Shared first author

Additional publications:

- 5) Galindez G†, **Sadegh S**†, Baumbach J, Kacprowski T, List M (2023). “Network-based approaches for modeling disease regulation and progression.” *Computational and Structural Biotechnology Journal*, 21, 780–795; <https://doi.org/10.1016/j.csbj.2022.12.022>
- 6) Bernett J, Krupke D, **Sadegh S**, Baumbach J, Fekete SP, Kacprowski T, List M, Blumenthal DB (2022). “Robust disease module mining via enumeration of diverse prize-collecting Steiner trees.” *Bioinformatics*, 38, 1600–1606; <https://doi.org/10.1093/bioinformatics/btab876>
- 7) Matschinske J†, Salgado-Albarrán M†, **Sadegh S**†, Bongiovanni D, Baumbach J, Blumenthal DB (2020). “Individuating Possibly Repurposable Drugs and Drug Targets for COVID-19 Treatment Through Hypothesis-

Driven Systems Medicine Using CoVex.” *Assay and Drug Development Technologies*, 18: 348–355; <https://doi.org/10.1089/adt.2020.1010>

- 8) Elbatreek MH, **Sadegh S**, Anastasi E, Guney E, Nogales C, Kacprowski T, Hassan AA, Teubner A, Huang PH, Hsu CY, Chiffers PMH, Janssen GM, Kleikers PWM, Wipat A, Baumbach J, Mey JGRD, Schmidt HHHW (2020). “NOX5-induced uncoupling of endothelial NO synthase is a causal mechanism and therapeutic target of an age-related hypertension endotype.” *PLoS Biology*, 18(11), e3000885; <https://doi.org/10.1371/journal.pbio.3000885>
- 9) Franziska Hufsky et al. (2021). “Computational strategies to combat COVID-19: useful tools to accelerate SARS-CoV-2 and coronavirus research.” *Briefings in Bioinformatics*, 22(2):642–663; <https://doi.org/10.1093/bib/bbaa232>
- 10) Nogales C, Grønning AGB, **Sadegh S**, Baumbach J, Schmidt HHHW (2021). “Network Medicine-Based Unbiased Disease Modules for Drug and Diagnostic Target Identification in ROSopathies.” *Handbook of experimental pharmacology*, 264:49–68; https://doi.org/10.1007/164_2020_386

† Shared first author

Table of Content

Acknowledgements	ii
Abstract	iii
Kurzfassung	v
Publication Record	vii
1. General Introduction	1
1.1. Motivation and Objective	1
1.2. Outline	4
2. Background	6
2.1. Central dogma of molecular biology	6
2.1.1. Molecular profiling	8
2.1.1.1. Genomics	8
2.1.1.2. Transcriptomics	9
2.1.1.3. Proteomics	10
2.2. Drug repurposing	12
2.3. Computational approaches for drug repurposing	15
2.4. Network and systems medicine	20
2.4.1. Molecular and biomedical networks	22
2.4.1.1. Network modeling and terminology	22
2.4.1.2. PPI networks	24
2.4.1.3. SARS-CoV-2 Virus-host interactome	25
2.4.1.4. Diseasomes	25
2.4.1.5. Comorbiditome	27
2.4.1.6. Drugome	27
2.4.2. De novo vs. traditional enrichment methods	27
2.4.3. Disease ontologies and vocabularies	30
2.4.4. Bird's-eye-view vs. close-up network medicine	32
2.5. Data integration for in silico and network-based drug repurposing	32
2.5.1. Challenges	32
2.5.2. Database models	34
3. General Methods	36
3.1. Overview of the drug repurposing platforms	36
3.2. Data integration and network construction	37

3.3.	Network medicine algorithms	39
3.4.	Testing the bias introduced to network medicine studies due to inadequate disease definitions	42
4.	Publications	45
4.1.	Publication 1: Exploring the SARS-CoV-2 virus-host-drug interactome for drug repurposing	45
4.2.	Publication 2: Network medicine for disease module identification and drug repurposing with the NeDRex platform	48
4.3.	Publication 3: Lessons from the COVID-19 pandemic for advancing computational drug repurposing strategies	51
4.4.	Publication 4: Lacking mechanistic disease definitions and corresponding association data hamper progress in network medicine and beyond	54
5.	General Discussion and Outlook	57
5.1.	COVID-19 drug repurposing with CoVex	58
5.2.	Interactive and integrative drug repurposing with NeDRex	59
5.3.	Lessons from COVID-19 to improve computational drug repurposing strategies	60
5.4.	The bias introduced to network medicine by inadequate disease definitions	63
5.5.	Limitations and challenges	66
5.6.	Conclusion and outlook	69
A.	Appendix	71
A.1.	Exploring the SARS-CoV-2 virus-host-drug interactome for drug repurposing	71
A.2.	Network medicine for disease module identification and drug repurposing with the NeDRex platform	81
A.3.	Lessons from the COVID-19 pandemic for advancing computational drug repurposing strategies	94
A.4.	Lacking mechanistic disease definitions and corresponding association data hamper progress in network medicine and beyond	104
	Acronyms	120
	References	122

1. General Introduction

1.1. Motivation and Objective

Advances in high-throughput omics technologies have enabled researchers to simultaneously analyze thousands of Deoxyribonucleic acid (DNA), Ribonucleic acid (RNA), proteins, and other molecules to gain a comprehensive and systemic insight into the inner workings of cells. This has led to the identification of genetic risk factors and molecular mechanisms involved in diseases through large-scale whole-genome and whole-transcriptome association studies. The availability of data from different molecular levels, including genomics, transcriptomics, and proteomics, provides opportunities to improve human health by offering safer treatments and making medicine more precise and personalized [1]. However, molecular profiling data is complex and often contains noise, making it difficult to determine which molecular changes cause a disease and which are caused by it. Researchers have attempted to deal with this complexity by incorporating prior information relevant to the whole population, such as protein-protein interactions networks, biological pathways, and ontological information, into their analyses [2]. This has allowed for integrative analyses, such as de novo network enrichment, to be performed. While prior information can be accessed for free through publicly available databases, obtaining multi-omics data for the same group of patients can be costly, hence, not as common in practice.

Systems medicine views the human body as a whole system rather than reducing it to the sum of its components. It uses the complex molecular interactions within the body to understand and explain diseases and their subtypes [3]. By focusing on the underlying mechanisms of disease, rather than just the symptoms, systems medicine can identify potential targets for treatment rather than just palliating symptoms. The identification of molecular entities that are connected in a way that helps to explain a disease phenotype is often called *disease module identification*. Many systems medicine approaches employ omics data together with the prior knowledge such as protein-protein interactions to derive disease modules. A *disease module* can include a variety of molecular entities, such as genes, proteins, and metabolites. The connectivity characteristic of a disease module implicates the molecular mechanism explaining the underlying pathological process [4]. Network medicine, a branch of systems medicine, employs network theory to represent biological systems as networks of interconnected nodes, where nodes symbolize biological elements and links represent their relationships [5]. This approach has multiple potential applications, including the identification of disease genes and pathways, which in turn may offer better targets for drug development and the development of better biomarkers.

A powerful strategy within drug development is drug repurposing, or drug repositioning, which involves identifying new applications for existing drugs beyond their primary medical indications. This approach offers several advantages over traditional de novo drug discovery, including a substantial reduction in development time and costs [6]. Repurposed drugs can typically reach the market in half the time and at a fraction of the cost compared to entirely new drug candidates [7]. Furthermore, the lower risk of failure due to safety concerns makes drug repurposing an attractive strategy.

In pandemic cases, like the recent coronavirus disease 2019 (COVID-19) pandemic, where quick response to a new unknown pathogen is vital, de novo drug discovery, due to its very long development timeline, is out of the question. For such cases, the best possible course of action to find treatment is drug repurposing [8]. Computational methods for drug repurposing, particularly those based on network medicine principles, can dramatically expedite the drug discovery process by narrowing down the search space for potential therapeutics.

Amongst diverse existing computational drug repurposing approaches, the network medicine approaches have gained high attention [9]. This is due to the fact that the nature of complex biological systems, associations between their constituent elements, as well as links between drugs and disease related components can be viewed in a holistic manner as networks by incorporating a variety of disease- and drug-related data types. By combining the principles of network medicine with large-scale biomedical data, such as disease-gene, drug-indication, drug-target, and disease-symptom relationships, it is possible to build meaningful models and extract valuable insights about diseases and drugs at the network level [10]. Many scattered biomedical databases contain information applicable to drug repurposing and they use non-unified identifiers for drugs and different vocabularies for diseases. However, to fully exploit such wealth of biological and pharmacological information for drug repurposing, an integrated and harmonized knowledge base is needed.

A large number of computational drug repurposing methods are developed based on machine learning principles to predict de novo drug-disease links. However, most of such methods work as a black box and their results are hardly interpretable. Understanding the rationale behind a model's predictions, is crucial for its clinical applicability. To address this limitation, a mechanism-based drug repurposing approach is emerging as a promising solution. This approach entails first identifying the disease mechanism through disease module mining and subsequently attempting to find drugs that target the mechanism. Such a mechanism-based drug repurposing approach returns interpretable results by design. Drugs targeting disease modules hold the potential to be disease modifying rather than just alleviating symptoms. Additionally, the valuable expert knowledge is often overlooked in computational drug repurposing methods.

Capitalizing on the idea of human-in-the-loop for drug repurposing enables us to evaluate the results of each step in a workflow by an expert leading to more promising predictions.

Available drug repurposing studies are limited to either non-translational algorithmic approaches or predictions for specific diseases. There are a few drug repurposing platforms attempting to make their methods applicable to more than just a limited set of diseases. However, their applicability is necessarily confined to only those diseases with extensive curated knowledge in public databases and thus renders these platforms impractical for new emerging diseases like the recent COVID-19 pandemic. Even for diseases with available knowledge in the databases, the analysis can be biased due to the inadequate, mainly phenotype-based disease definitions and annotations. In other words, patients diagnosed with the same (symptom-based) disease, do not necessarily imply the same disease mechanism. Therefore, the drug repurposing field is in need of integrated and interactive tools employing mechanistic drug repurposing methods and allowing biomedical researchers to employ network-based drug repurposing approaches that are adaptable to their individual use cases while also exploiting their expert knowledge. The main objective of this dissertation is to address this need by developing integrative and interactive network medicine platforms for biomedical researchers. My secondary goal is to assess a less explored type of bias that disease-associated data introduces to network medicine approaches, originating from inadequate disease definitions.

In the first publication [11], I developed CoVex (CoronaVirus Explorer), a systems and network medicine platform to facilitate the interactive integration of expert knowledge, such as knowledge about virus replication, immune-related biological processes, virus pathomechanism, or drug mechanisms, on top of the publicly available related data with the aim of drug repurposing for the COVID-19 disease [11]. The platform is accessible via a user-friendly web interface. CoVex integrates experimentally validated virus-human interaction data for SARS-CoV-2 and SARS-CoV-1, experimentally validated PPIs in humans, as well as drug information including drug-target and clinical trial data. It uses this wealth of data jointly with network medicine algorithms including my novel Multi-Steiner Tree (MuST) method to predict novel drug targets and drug candidates. Furthermore, it allows researchers to use their clinical data as a starting point and to test their hypothesis augmented by exploring the molecular mechanisms visually.

In the second publication [12], I present the NeDRex platform which generalizes the approach implemented in CoVex beyond COVID-19 drug repurposing to be applicable for other diseases. It allows to construct heterogeneous biological networks enriched with disease and drug data, to mine the networks for candidate disease mechanisms while benefiting from the interactive engagement of expert knowledge, and finally, to prioritize drugs directly or indirectly targeting proteins

involved in the mechanisms. NeDRex is composed of three main components: a knowledge base (NeDRexDB, accessible via a RESTful API and a Neo4j endpoint), an API (NeDRexAPI), and a Cytoscape app (NeDRexApp) which is the main user interface of the platform. I developed a statistical method for the validation of predicted disease mechanisms and drug candidates as part of the platform. In this publication, I also demonstrated the utility of the platform by showcasing five different diseases including ovarian cancer, inflammatory bowel disease, pulmonary embolism, Huntington's disease, and Alzheimer's disease.

My third publication [13] is a comprehensive review of computational approaches for drug repurposing in the context of COVID-19, divided into virus-targeting and host-targeting approaches, accompanied by the introduction of the most relevant data resources. This review also reflects on the knowledge gained from the studies analyzed and proposes a unified drug repurposing strategy to enhance preparedness for probable future outbreaks. This unified strategy is not exclusive to COVID-19 drug repurposing or urgent pandemic cases but also applicable to general drug repurposing.

My fourth publication addresses an often ignored, nevertheless, important source of bias for any disease-involved analysis: the influence of ill-classified disease definitions. Lumping symptomatically similar, but mechanistically different diseases together under one disease name will blur the signal and hinder meaningful discovery. This problem is prevalent in all domains of biomedical research and with this study we were finally able to quantify the bias arising through these fuzzy disease definitions, specifically within the domain of network medicine approaches that exploit large-scale disease-related data from public databases.

I have published numerous other papers during my PhD study which are not fully discussed in the dissertation but nevertheless are relevant. I contributed to a publication where we used the hypothesis-driven method from the CoVex tool to explore potential repurposable drugs for COVID-19 (joint first-authorship) [14]. I also contributed to another publication where we improved our disease module identification method MuST, that was implemented for CoVex and NeDRex, to be more robust [15]. Finally, I contributed to a review work on network-based approaches for modeling disease regulation and progression (joint first-authorship) [16]. The full list of published papers is given in the Publication Record section.

1.2. Outline

The Background chapter begins with elucidating fundamental concepts of molecular biology (2.1), followed by an introduction to the concept of drug repurposing (2.2). Subsequently, it provides an overview of different types of

computational approaches employed in drug repurposing (2.3). Delving into the network and systems medicine discipline (2.4), this section details various types of networks central to this dissertation (2.4.1), presents various types of disease module identification methods (2.4.2), and explains the foundation of how diseases are currently defined (2.4.3). Finally, it explains data integration for in silico and network-based drug repurposing together with its associated challenges (2.5).

The General Methods chapter provides an overview of the drug repurposing platforms and their components, forming the foundation of this dissertation (3.1). Then the data integration and network construction tasks that underlie publications 1, 2 and 4 are touched upon (3.2). Afterwards network medicine algorithms that were adapted or de novo developed for the purpose of disease module identification and drug ranking as part of the CoVex and NeDRex platforms are briefly introduced (3.3). Lastly, the framework employed to test the bias introduced to network medicine studies by inadequate disease definitions is described (3.4).

Chapter 4 provides the summaries of all four publications included in this dissertation, along with an explanation of my contribution. The complete versions of the papers are incorporated in the Appendix section.

In the General Discussion and Outlook chapter (5), I address both the limitations of our work and the broader challenges faced by the field. I propose various strategies to address the outlined constraints. Finally, I explore the future prospects of the drug repurposing domain and draw conclusions regarding the conducted Ph.D. project.

2. Background

2.1. Central dogma of molecular biology

The central dogma of molecular biology is the fundamental process in which genetic information flows from DNA to RNA (by transcription) and then becomes a functional protein product (by translation). This concept was originally proposed by Francis Crick in 1957 [17], who believed that genetic information could not be transferred back to DNA once it had been turned into protein. Although our understanding of genetic information has advanced significantly over the past decades, Crick's original idea has remained important and is commonly used to denote the general way in which sequence information is exchanged within cells.

The connections between DNA, RNA and proteins are highly complex, and while the fundamental principles of the central dogma apply to both eukaryotes and prokaryotes, the details of these processes exhibit notable variations. Replication, transcription, and translation are the three fundamental systems that maintain the interchange of genetic information [18] (Fig. 2.1). The process by which DNA is duplicated during cell division is known as *DNA replication*. DNA helicase splits the two complementary DNA strands apart during replication, creating a replication fork. Each strand serves as a template for DNA polymerase proteins to construct a new two-stranded DNA molecule. A DNA transcription unit is copied into an RNA molecule during *transcription*. A coding sequence and several regulatory components, including promoters, make up a transcription unit. RNA polymerase and one or more transcription factors that have bound to the promoter initiate transcription. After creating a transcription bubble by separating the DNA strands, RNA polymerase adds RNA nucleotides to the template strand to produce an RNA complement. In eukaryotes, the generated pre-RNA goes through additional processing, such as splicing out introns, to create messenger RNA (mRNA). The mRNA serves as a template for the synthesis of a polypeptide during *translation*. As it travels along the mRNA strand, a ribosome reads nucleotide triplets. Beginning at the start codon and continuing until a stop codon is reached, a peptide chain is built from amino acids that match the triplet sequence. The protein (polypeptide) will fold into a three-dimensional (3D) structure during assembly, which will dictate its function.

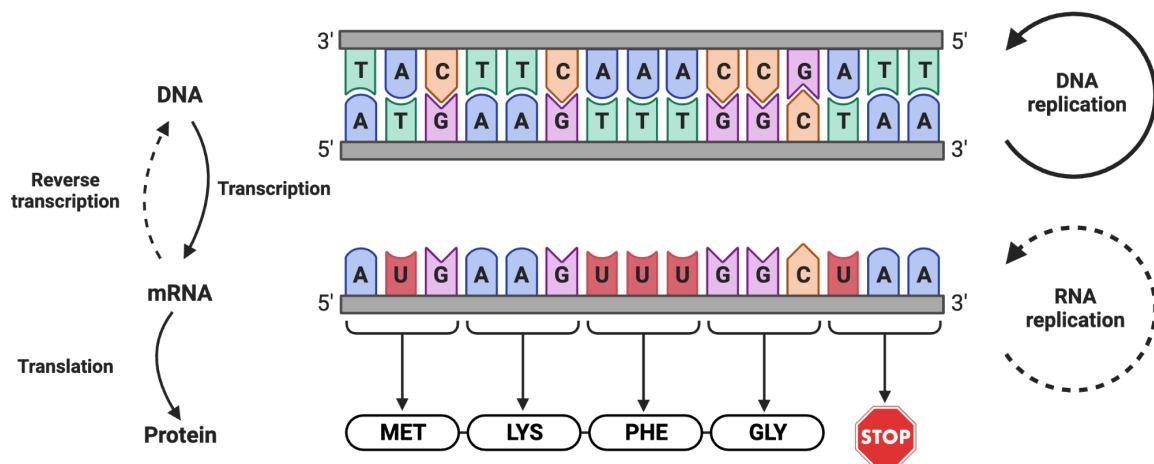


Figure 2.1 - The central dogma of molecular biology. Genetic information flows from DNA to mRNA through transcription and then from mRNA to proteins through translation. Unusual flows of information marked in dashed lines. Created with BioRender.com

As many additional processes participate in the regulation of genetic information, it is now clear that the central dogma offers an oversimplified understanding of gene expression. There are some cases of genetic information flow that are not in the original dogma, including reverse transcription, a mechanism by which genetic information is transferred from RNA to DNA (occurring in the case of retroviruses, as well as in eukaryotes, e.g., in the telomere synthesis) and RNA replication, the process of copying of one RNA to another. Gene expression can also be profoundly impacted by epigenetic modifications including DNA methylation and histone modifications [19]. Additionally, not every transcription unit codes for a protein; some instead produce regulatory non-coding RNAs like micro RNAs (miRNA) [20]. Regardless of this, the central dogma of molecular biology remains at the center of bioinformatics research. Understanding cells at the DNA, RNA, and protein levels has been essential to unraveling cellular function and has helped molecular biologists gain a deeper understanding of the inner workings of living organisms, specifically how changes at the molecular level can significantly impact the overall system and affect the phenotype. One example of this is mutations in DNA, which can disrupt the flow of information. Some mutations have no known impact on the protein sequence, although others have a significant impact on protein function by allowing proteins to lose or gain functionality [21]. Monogenic disorders are caused by mutation in a single gene and are mostly identified by their distinctive patterns of familial inheritance. Examples of this type of disorder include Huntington disease, sickle cell anemia, cystic fibrosis, and Duchenne muscular dystrophy. Dr. Victor A. McKusick founded the database Online Mendelian Inheritance in Man (OMIM) in 1997 with a focus on inherited genetic disorders in humans [22]. It is periodically updated and as of July 14, 2023, OMIM reported 4,813 genes with a phenotype-causing mutation, and 7,389

phenotypes with a known molecular basis (<http://omim.org/statistics/geneMap>).

However, most diseases are not caused by a single gene mutation but rather a combination of genetic variations occurring in several locations as well as environmental factors. Some examples of complex disorders include cancer, asthma, Alzheimer's disease, Parkinson's disease, multiple sclerosis, osteoporosis, kidney diseases, autoimmune diseases, and many more [23].

2.1.1. Molecular profiling

2.1.1.1. Genomics

Genomics is the study of an organism's complete set of DNA – called the genome. Early genomics research focused on sequencing the complete genomes. The first comprehensive euchromatic human genome was published by the Human Genome Project in 2004 [24]. The genome of a person, consisting of 3.2 billion base pairs, is roughly 99.8% similar to the genomes of all other humans. The variations in the remaining 0.2% (4-5 million sites) are of paramount importance for understanding the differences between healthy and disease conditions allowing researchers to systematically investigate the causes of disease. Genomic variations can be grouped into two types: single-nucleotide variations (SNVs), including single-nucleotide polymorphisms (SNPs) and small indels (insertions and deletions), and structural variations (SVs), including but not limited to inversions and copy number variations (CNVs) [25]. Variations may happen in coding, non-coding, or intergenic regions of the genome. SNPs in a coding region do not necessarily change the amino acid sequence of the protein that is produced (synonymous) but these SNPs still can affect the function of proteins in other ways. When a SNP affects the protein sequence (nonsynonymous), two scenarios can happen: missense point mutation where one amino acid in a protein changes; or nonsense mutation which results in a premature stop codon that could lead to a nonfunctional protein product. SNPs that are not in protein-coding regions can be consequential and still affect gene splicing, transcription factor binding, and mRNA degradation [26].

Rapid advances in high-throughput sequencing technology at the beginning of the 21st century have rapidly reduced the cost and time of sequencing, making it possible to study the genomes of a large number of individuals. Even before high-throughput sequencing, DNA microarray technology paved the way to the development of Genome-Wide Association Studies (GWAS), which seek to identify associations between particular traits or conditions and variations at a single position in DNA (SNPs). The goal of GWAS is to create a comprehensive list of SNP-condition associations, making the relationship between genotype and phenotype relatively simple. However, GWAS has limitations, including difficulty in studying

rare variants and statistical issues when testing for millions of SNPs. With next-generation sequencing (NGS), DNA sequencing has emerged as a promising alternative to array-based techniques. Whole-genome sequencing (WGS) avoids the bias from probe (used to selectively target and analyze specific regions of the genome) selection and is able to detect rare variants. Exome sequencing has also been developed as a more efficient and cheaper alternative to WGS by focusing specifically on protein-coding DNA [27]. CNVs are structural variations where a large segment (e.g., greater than 1 kilobases) of the genome is duplicated or deleted [28]. Although a typical genome contains relatively few CNVs, according to the CNV map, 4.8–9.7% of the human genome is involved with CNVs [29]. WGS-based methods are able to detect CNVs. They can differ greatly among individuals and can also play a role in many genetic disorders [30].

2.1.1.2. Transcriptomics

The study of transcriptomes, or the complete set of RNA molecules present in a cell, is known as transcriptomics. Whereas genomics analysis looks at what the cell is able to do, gene expression analysis gives us information about what the cell is currently doing. Although measuring the actual abundance of proteins would give a more direct estimate of protein activity, proteomics analysis is complex, partly, because many different proteins can be produced from a single gene due to alternative splicing (AS), SNPs and post-translational modifications [31]. This makes gene expression a tempting proxy for protein activity. However, some studies have reported the correlation between mRNA and protein abundance to be rather modest, suggesting gene expression may be misleading for measuring protein activity under some conditions [32–34].

Similar to genomics, microarrays have been the dominant method for gene expression analysis until recently. However, in recent years, RNA-seq (or whole-transcriptome shotgun sequencing) has emerged as the primary approach for analyzing the entire transcriptome. This shift is attributed to the significant reduction in sequencing costs and advancements in computational methods. [35]. Large amounts of gene expression data is already publicly available in data repositories, such as the Gene Expression Omnibus (GEO) [36], and through large-scale research initiatives like the Genomic Data Commons [37].

One of the main applications of gene expression data is differential expression analysis, in which the expression level of each gene is tested for association with a phenotype (e.g., disease status) using an appropriate test statistic. While many early studies have used two-sample hypothesis tests, more modern gene expression analysis methods are based on linear models, like DESeq2 [38] and limma [39]. Gene expression data is also used in unsupervised analysis to discover genes that are functionally related [40], to separate patients into clinically

relevant subgroups [41–43], or to simultaneously identify molecular mechanisms responsible for the patient grouping [44].

AS, a process allowing a single gene to code for multiple proteins [45] happens in over 90% of human genes and is a major mechanism for the diversification of transcriptome and proteome [45]. In addition to playing an important role in normal cellular processes, a variety of pathogenic processes underlying many different diseases involve AS [46]. Due to the potential confounding effect of gene expression levels and the often-limited number of patients with relevant RNA-seq data, the robust identification of disease-associated splice events based on RNA-seq data is challenging [46].

2.1.1.3. Proteomics

Proteins are functional units of cells and essential for gene regulation and functioning of the entire body. During the translation process of mRNA to protein, the molecular complexity increases by several mechanisms. Alternative splicing of the transcriptome generates between 70K and 100K transcripts from the initial 20K human genes [47]. These transcripts are translated into an even greater variety of unique protein sequences due to occurrence of sequence mutations [48] and alternative translation [49]. The expressed proteins assemble themselves into complexes [50,51], may carry post-translational modifications, have different subcellular localizations [52] and are differentially degraded [53,54]. This leads to a high complexity in the proteome, where individual proteins can be found in a range of abundances from a few hundred molecules for gene regulatory proteins to millions of molecules for structural proteins.

A disease can be characterized as the result of imbalanced information flow in a biological system leading to an altered proteome [55]. The field of proteomics is formed to study the complete set of proteins (proteome) in an organism with the goal of linking changes in the proteome to various phenotypes and diseases.

The lack of a method to amplify individual proteins before detection, unlike in genomics, requires protein analytics to have high sensitivity to be able to detect low abundances of individual analytes in complex protein mixtures and handle the variation in their abundance levels (dynamic range) within a sample. Mass spectrometry (MS) involves determining the mass-to-charge ratio of ions and presenting it as a graph of intensity, known as a mass spectrum. Current approaches developed based on MS allow for high-throughput proteome analysis. They manage the dynamic range of protein expression levels, quantify amounts of analytes from complex mixtures, and enable comparisons of numerous biological samples to identify variations in their protein profiles with high sensitivity [56].

Other methods include chromatography-based (also in conjunction with MS), enzyme-linked immunosorbent assay (ELISA) for selective protein analysis, antibody-based affinity methods for tracking the expression of specific proteins

in tissues, protein microarrays or chips for high-throughput and quick expression analysis, and 2D-gel-based approaches for separating complex protein samples [56].

A wide range of scientific topics are addressed by proteomics analysis, including (i) determining the molecular function of a protein, (ii) finding links between disease and variations in protein structures, (iii) protein-protein interactions (PPIs), (iv) drug discovery, and many more.

Protein Interactions - In addition to protein sequence and structure, expression profile, post-translational modifications, and intracellular localization, the function of a protein is influenced by interactions with other proteins [57]. Most cellular processes are not maintained by a single protein but rather by complexes of multiple proteins linked by PPIs.

These interactions are established through highly specific physical contacts. Based on the type of interacting partners, stability of the PPI complexes, and nature of the interface between the proteins, the PPIs can be classified into two categories [58,59]. Stable interactions form stable protein complexes, whereas transient interactions play a role in cellular processes, such as protein modification, transportation, folding, signaling, and cell cycling [57].

A wide range of experimental and prediction-based methods have been developed for PPI detection. Tandem Affinity Purification-MS (TAP-MS) and Yeast Two-Hybrid (Y2H) screening are two widely used in vitro methods for high-throughput PPI detection. TAP-MS is capable of identifying both individual protein interactions and multi-protein complexes with high accuracy, but it can only detect stable interactions and misses short-lived ones due to the tandem purification steps. On the other hand, Y2H screening is a simpler and cost-effective method for detecting PPIs in vivo, including transient interactions, but has a high rate of false positives. The AP methods may overlook low-abundance proteins, while Y2H does not have this issue [60]. Co-immunoprecipitation is considered to be the best method for identifying PPIs, but it has limitations. It requires the selection of an antibody that specifically targets a protein believed to be involved in an interaction, making it unsuitable for large-scale screening. However, it is highly effective in verifying interactions that have been identified through high-throughput techniques [61]. PPIs have also been predicted in silico using various features of proteins, such as protein sequence, 3D structure, co-expression, network topology and Gene Ontology (GO) annotations [62–64].

By investigating protein interactions within cells and biological systems, researchers gain valuable insights into disease mechanisms and potential therapeutic targets. A very good example of such studies is the study of Gordon et al. [65] that helped researchers early during the recent COVID-19 pandemic gain valuable knowledge about the Severe Acute Respiratory Syndrome Coronavirus 2

(SARS-CoV-2) and its interaction with human proteome. Using affinity-purification (AP) MS, they could identify 332 high-confidence PPIs between SARS-CoV-2 and human proteins. The human-virus interactome from this study became the cornerstone of many subsequent studies for identifying drug targets.

2.2. Drug repurposing

Over the past couple of decades, there have been significant advances in the scientific and technological fields contributing to drug research and development (R&D). To name a few, there has been a great increase in the size of chemical libraries owing to combinatorial chemistry, which increased the number of drug-like molecules that can be synthesized per chemist per year [66]. Faster DNA sequencing has led to the identification of new drug targets [67]. Faster 3D protein structure calculation via X-ray crystallography has facilitated the identification of improved lead compounds through structure-guided strategies [68]. High-throughput screening (HTS) has resulted in a major reduction in the cost of testing compound libraries against protein targets [69]. Last but not least, computational drug design and screening have contributed to advances in scientific knowledge, including understanding of disease mechanisms, new drug targets, biomarkers and surrogate end points [66].

However, these advances have not led to an increase in the number of new approved drugs per R&D spendings in the drug industry and the efficiency of R&D of new drugs has decreased constantly (Fig. 2.2). This trend is called “Eroom's Law”, in contrast to the familiar Moore's Law ('Eroom's Law' is 'Moore's Law' in reverse). This dramatic decline is opposite to the IT industry, where Moore's Law described the exponential increase in the number of transistors that can be placed at a reasonable cost onto an integrated circuit. This number doubled approximately every two years from the 1970s. While, the number of new drugs introduced per year has been broadly flat over the period 1950s-2010s, and costs have grown fairly steadily [66]. Although this trend has changed over the past ten years [70], the cost of development of a new pharmacologically active drug, including the whole process from traditional drug discovery to market introduction, is still estimated between two and three billion USD [71]. Moreover, total development time using the de novo drug discovery approach, i.e., entirely developing a new drug for a given indication, takes 12-16 years [7]. Further, de novo drug discovery suffers from a high attrition rate, i.e., only 10% Of the drugs entering phase 1 clinical trials are approved, the rest fail because of high toxicity or inefficacy [7,72]. Therefore, there is an increasing interest in discovering alternative approaches for finding therapeutics.

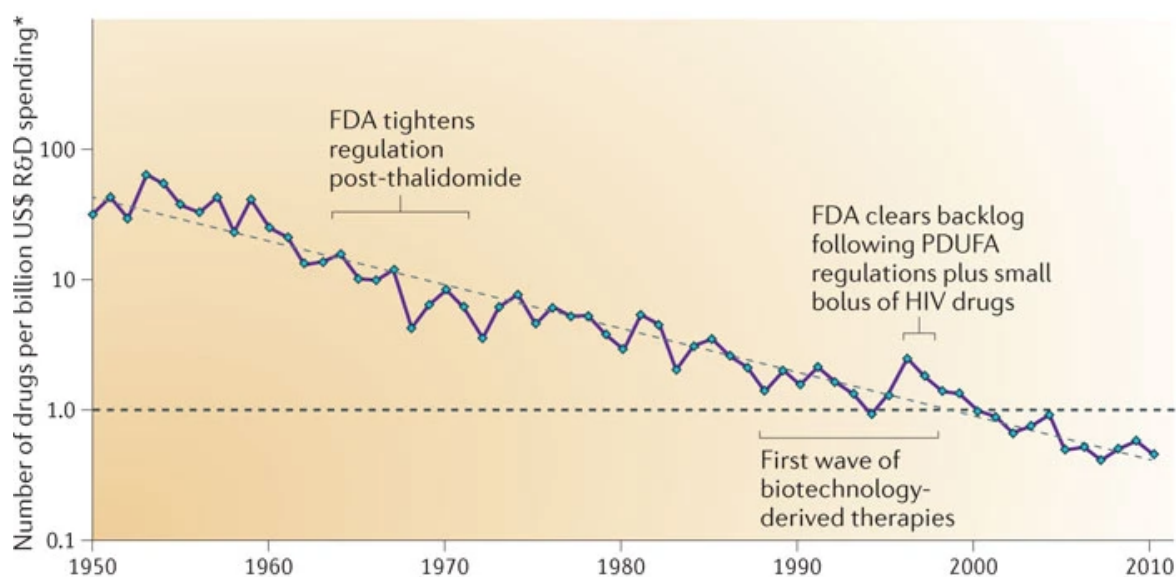


Figure 2.2 - Eroom's Law. The rate at which new drugs are approved by the US Food and Drug Administration (FDA) for every billion US dollars (adjusted for inflation) invested in research and development (R&D) has decreased by approximately 50% every 9 years. The figure is taken from Scannel et al. [66]. Permission granted by Springer Nature.

Drug repurposing (also known as drug repositioning) is the process of identifying alternative uses for approved or investigational drugs that are not in the scope of the primary medical indication [6]. The drug repurposing strategy tries to address the aforementioned challenges de novo drug discovery is facing. Since by employing repurposing strategies most of the preclinical testing, safety assessment, and formulation development will already have been completed, the drug development time frame and total necessary investment can be reduced [73]. Repurposed drugs are generally approved on average within 6 years, significantly shorter than de novo developed drugs [7]. There can be considerable savings in preclinical and phase I and II costs for a repurposed drug compared to a new drug. The total costs of introducing a repurposed drug to market have been estimated to be US\$300 million on average [74] (Fig. 2.3). More importantly, drug repurposing benefits from a lower risk of failure from the safety aspect. Since the repurposed drug has already passed required safety by being tested in preclinical models and humans in early-stage trials, it is less likely to fail in subsequent efficacy trials due to safety reasons [73].

One of the oldest examples of drug repurposing is acetylsalicylic acid (Aspirin). It was first marketed in 1899 as an analgesic, and later in the 1980s repositioned as an anti-platelet aggregation drug [75]. Another example of successful repurposed drugs is Zidovudine, originally indicated and approved for cancer in 1987 and later was approved by FDA as the first anti-HIV drug [71,72].

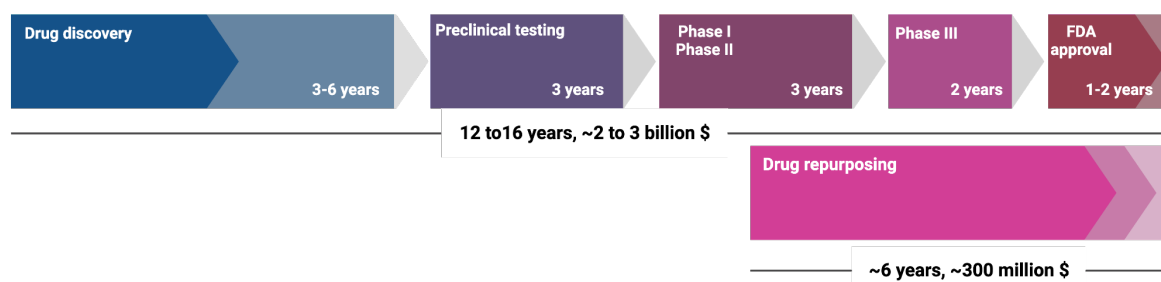


Figure 2.3 – Shorter timescale and reduced cost for repurposed drugs. Because most repurposed drugs have already passed the early phases of development and (pre-)clinical testing, they can potentially get approval in less time and at lower cost compared to de novo developed drugs. Figure modified from [74] and created with BioRender.com.

While early examples of successfully repurposed drugs have been largely serendipitous and most successful repurposing examples so far have not involved a systematic approach [6], advances in omics technologies and the availability of massive amounts of omics data have provided opportunities for systematic *in silico* prediction of new drug-disease relationships. In the case of a pandemic which requires fast reaction, like the recent example of COVID-19, there is an urgent need for systematic approaches which could significantly speed up drug discovery. An effective strategy involves narrowing down the search to existing drugs by using drug repurposing methods, which has the potential to significantly expedite the typically lengthy approval process. [13].

A standard drug repurposing pipeline consists of a three-step process before the candidate drug can enter the late phases of development pipeline: (1) identification of a candidate drug for a given indication (hypothesis generation); (2) mechanistic assessment of the drug effect in preclinical models; and (3) efficacy evaluation in phase II clinical trials. Systematic approaches for hypothesis generation could be most beneficial in the critical step of identification of the right drug for an indication of interest with a high level of confidence (step 1). These systematic approaches can be grouped into computational and experimental approaches (Fig. 2.4), both being more and more used synergistically [71].

The experimental approaches can be categorized into two groups: 1) Identifying relevant target interactions with binding assays 2) High throughput phenotypic screening of compounds using *in vitro* or *in vivo* disease models to indicate potential candidates [71]. Different categories of computational approaches are discussed more in detail in the subsequent section (2.3).

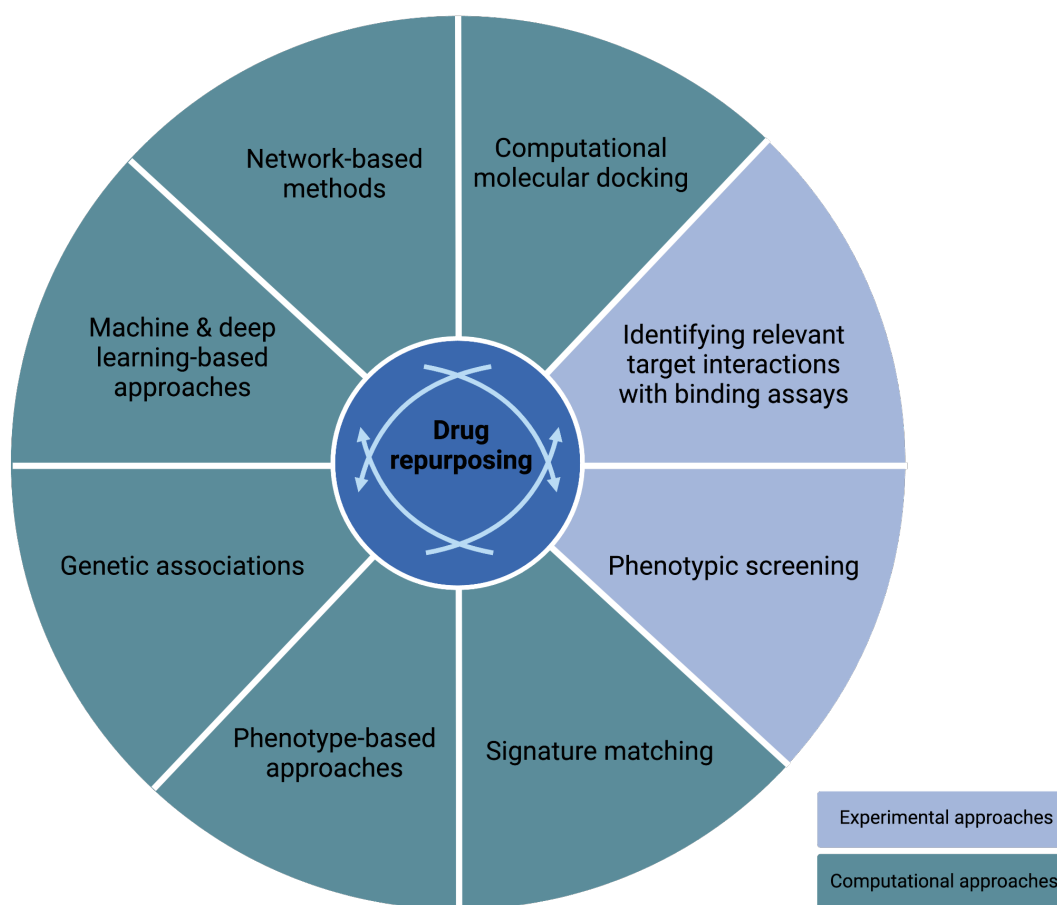


Figure 2.4 - Drug repurposing approaches. Different computational and experimental approaches can be used individually or jointly to systematically analyse different types of large-scale data to generate repurposing hypotheses and meaningfully interpret the repurposed candidates. Figure inspired by [71] featuring altered classifications and created with BioRender.com.

2.3. Computational approaches for drug repurposing

There are a variety of classifications for computational drug repurposing methods; each of which seeks to categorize the existing methods depending on some different important aspects. Here we consider the following adaptation of classification of computational drug repurposing methods mainly based on the work of Pushpakom et al. [71] and K. Park [72]:

Signature matching - Signature matching involves comparing the distinct characteristics or "signature" of one drug with that of another drug or disease [76,77]. Drug signatures can be derived from two primary types of data: omics data including transcriptomic (RNA), proteomic, and metabolomic data, as well as chemical structures. Matching transcriptomic signatures can be used to evaluate drug-disease similarity [78] and drug-drug similarity [79].

The drug-disease similarity approach utilizes the degree of negative correlation between the gene expression signature of a drug and that of a disease. This correlation signifies the reversal of gene expression patterns, where genes upregulated in the disease are downregulated by the drug, and vice versa. Based on this principle, it becomes possible to infer whether the drug might have a potential impact on the disease. The principle of signature reversion underlies this approach, which assumes that if a drug has the ability to reverse the expression pattern of a specific gene set that characterizes a particular disease phenotype (i.e., a drug signature is closer to that of a healthy state), then the drug has the potential to revert the disease phenotype [71] (Fig. 2.5.A). The drug-drug similarity approaches try to discover common mechanisms of action among drugs that may appear dissimilar, such as drugs from different classes or with dissimilar chemical structures. This principle, known as 'guilt-by-association' serves to identify alternative targets of existing drugs and unveil potential off-target effects that can be further explored for clinical applications [80]. Therefore, if two drugs exhibit a shared transcriptomic signature, it suggests that they might also have a common therapeutic application, irrespective of the similarity or dissimilarity in their chemical structures [81] (Fig. 2.5.B). Both drug-drug and drug-disease similarity approaches rely on transcriptomic signature matching which requires publicly available gene expression data. One of the most comprehensive sources is the Connectivity Map (cMap) [82] the latest version consisting of over 1.5M gene expression profiles from ~5,000 small-molecule compounds, and ~3,000 genetic reagents, tested in multiple cell types.

The second approach to signature matching involves examining chemical structures and their association with biological activity. By comparing the chemical signatures of different drugs, one can identify chemical similarities that may indicate shared biological activity [83].

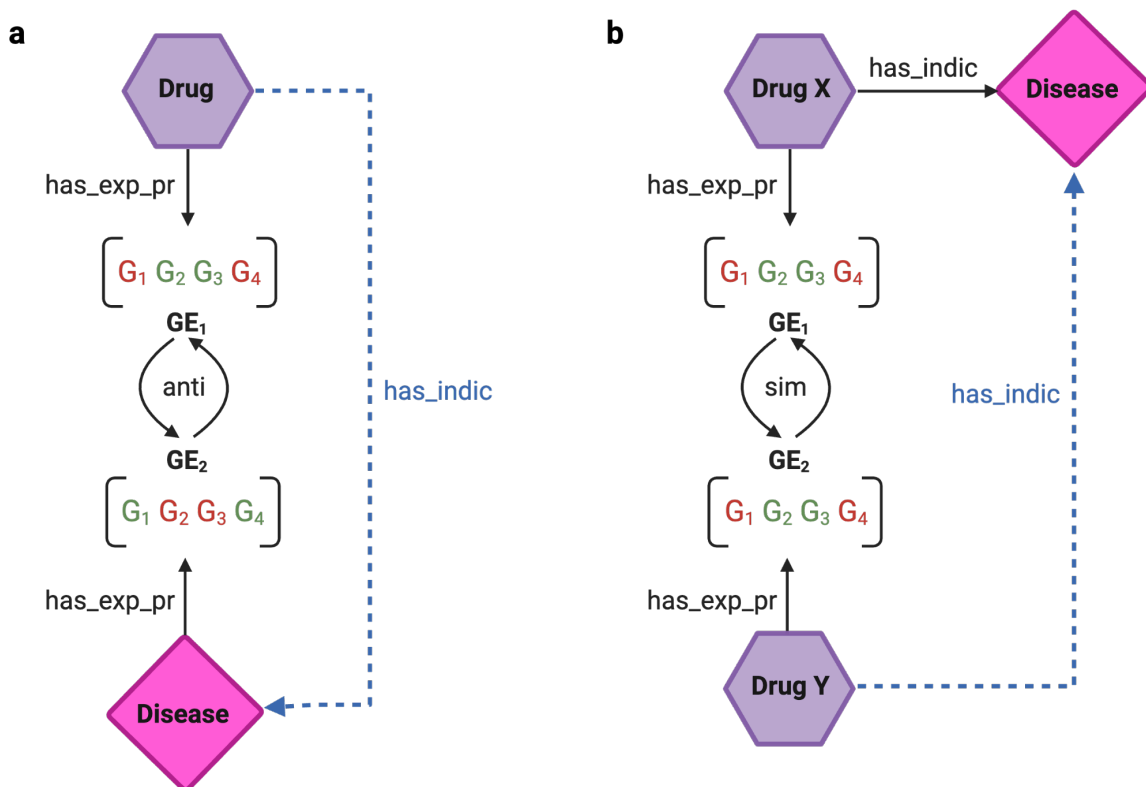


Figure 2.5 - Drug repurposing using signature matching. a) Drug-disease similarity approach. b) Drug-drug similarity approach. G = gene, GE = gene expression profile, *has_exp_pr* = has expression profile, *has_indic* = has indication, *anti* = anti-correlated expression profile, *sim* = similar expression profile. Blue dashed line represents an inferred indication. Created with BioRender.com

Computational molecular docking - Due to similarities in protein binding sites, drugs often have the ability to bind to "off-target" proteins. If it is known that an off-target protein is implicated in another disease, the drug holds potential for treating the second disease [84]. Computational molecular docking is an approach that employs structural information to predict the compatibility of a ligand (e.g., on the drug side) with a therapeutic target (e.g., a receptor on the protein side) by assessing their binding site complementarity, i.e., it predicts how two molecules interact in 3D space. When there is existing knowledge about a target associated with a disease, conventional docking can be used to evaluate multiple drugs against that specific target (one target and multiple ligands). On the other hand, inverse docking allows the exploration of drug libraries against various target receptors (several targets and one ligand), aiming to uncover novel interactions that hold potential for further repurposing efforts [71]. Molecular docking for drug repurposing faces challenges including limited availability of 3D structures for certain protein targets (particularly membrane proteins), the lack of well-curated databases with accurate structural information, although improving, and questionable reliability of docking algorithms in predicting binding affinity.

Genetic associations - There has been a large increase in the number of GWAS conducted over the past 15 years following advances made in genotyping technology. GWAS seek to identify genetic variants associated with diseases and consequently shed light on the biology of diseases. These associated variants can aid in the identification of new therapeutic targets. Some of them could be shared between multiple diseases already treated by drugs. Integrating disease associated genes data from GWAS with drug targets data available from public databases makes it possible to link drugs to disease traits which are not the same as the original indication of drugs [85] (Fig. 2.6).

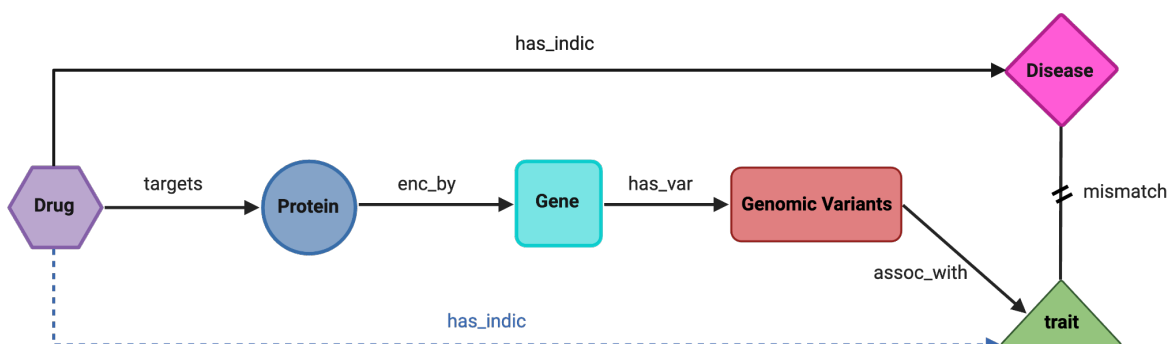


Figure 2.6 - Drug repurposing using genetic associations. An abstract view on using GWAS data for the discovery of potential drug repurposing candidates. *enc_by* = encoded by, *has_var* = has variants, *assoc_with* = associates with, *has_indic* = has indication, *mismatch* = not the same. Blue dashed line represents an inferred indication. Created with BioRender.com

While certain targets identified through GWAS or alternative methods may have the potential to be directly targeted by drugs, frequently these genes may not be ideal druggable targets e.g., the variants in non-coding regions. In such cases, a pathway-based strategy could seek for genes that are either upstream or downstream of the GWAS-associated target and could be druggable [86]. In a generalized fashion, a network-based strategy may seek to target interacting partners which could be used for repurposing [87].

Network-based methods - Network-based approaches hold promise, but there is still substantial progress to be made. Network-centric methods are particularly beneficial due to their inherent capacity to represent complex biological associations and offer a structured framework for integrating various data types and biological concepts and their interactions. In the models used by network-based approaches, network nodes represent drugs, proteins, or diseases, and edges indicate interactions or relationships between nodes, such as drug-drug similarities, drug-target interactions, gene-disease associations, and protein-protein interactions [88]. Networks can be knowledge-based (derived from direct evidence) or computationally inferred from multiple data sources. Some network-based methods use the 'guilt-by-association' principle in their heterogeneous molecular network to discover unknown drug-disease relationships [72]. Many

drug repurposing pipelines use a combination of computational approaches. For instance, some of the signature matching studies also employ the network analysis approach [81,89,90]. Many of the approaches described in this section, at some point in their workflow, use networks for the presentation of their underlying data and/or apply network-based algorithms to their data. The approach we used for developing the drug repurposing NeDRex platform (my second publication) can be considered a hybrid approach combining both genetic association and network analyses such that input seeds for the network-based disease module detection algorithms can be obtained from gene association studies [12].

Phenotype-based approaches - Phenotype-based strategies for drug repurposing focus on diseases or side-effects. These approaches involve connecting diseases based on shared characteristics, such as the underlying cause of the pathology or the observed biological dysfunction. Approaches that construct networks of diseases based on the similarity between them try to create a comprehensive "diseasome" perspective [91]. Section 2.4.1.4 elaborates more on the notion of diseasome. For instance, Li et al. [92] employed data pertaining to disease-associated genes and diseases-associated pathways. They then established connections between diseases based on shared pathways, hypothesizing that diseases exhibiting common dysregulated pathways demonstrated similarity. Their objective was to uncover new relationships among diseases, with the aim of facilitating pathway-guided therapeutic interventions for various conditions. Such studies use network presentation of the data and can also be considered a hybrid approach combining both phenotype-based and network-based approaches.

Side effects of drugs result from off-target activity of the drug, i.e., the biological activity of a drug which is different from its intended biological target. Studying side effects of drugs has the potential of discovering novel therapeutic uses for drugs. Some studies like Yang et al. [93] and Ye et al. [94] are grounded on the hypothesis that if two drugs induce similar side effects, they might be acting on a shared target, protein, or pathway [78]. In this context, side effects can serve as indicators of a shared underlying mechanism of action, and when two drugs have similar profiles of side effects, they can potentially be used in treating the same pathology (Fig. 2.7). Such studies use side effect information from databases like SIDER. Furthermore, it is plausible for the adverse effect phenotype of a specific drug to resemble that of a disease, implying shared pathways and physiological characteristics between the drug and the disease.

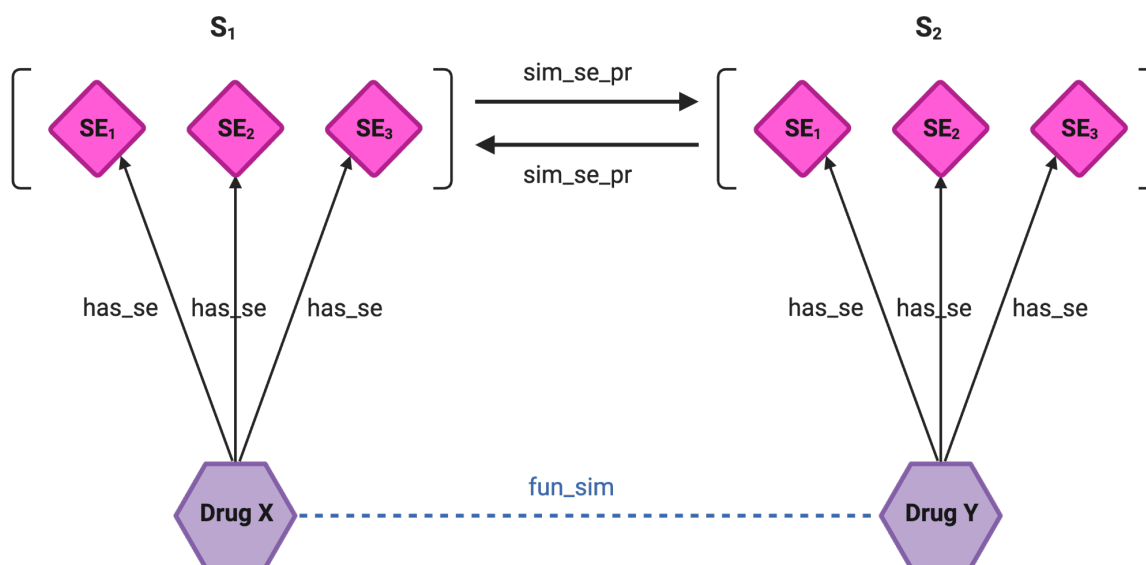


Figure 2.7 - Drug repurposing using side-effect similarity. An abstract view on using side-effect similarity profiles to infer potential drug repurposing candidates. *SE* = side-effect, *S* = set of side effects associated with drug, *sim_se_pr* = similar side-effect profiles, *has_se* = has side-effect, *fun_sim* = inferred functional similarity. Created with BioRender.com

Machine & deep learning-based approaches - A typical deep learning-based drug repurposing pipeline includes four steps: (1) integrate data sources enriched with information on drugs, proteins, and diseases; (2) generate informative feature vectors using various representation approaches (such as graphs, sequences, and text); (3) build and evaluate a deep learning model; and (4) conduct drug repurposing tasks, including prediction of drug–target binding affinity, drug–target interaction, compound–protein interaction, and drug–disease associations [95]. Aliper et al. showed that deep neural networks trained on large transcriptional response data sets from the LINCS Project could classify various drugs to therapeutic categories [96].

Computational prediction of binding affinity between compounds and targets greatly enhances the probability of finding lead compounds by reducing the number of wet-lab experiments used in experimental approaches. To improve drug–target interaction prediction in virtual screening, machine-learning techniques have been increasingly employed for predicting binding affinities using ligand-based and target-based approaches [97].

2.4. Network and systems medicine

Systems medicine represents an evolving field of study that perceives the human body as a whole system rather than reducing it to the sum of its components [3]. This interdisciplinary domain is characterized by its rapid evolution over time and focuses on understanding how cellular and tissue interactions give rise to physiological functions, shaping the behavior of these components in the human

body. Its conceptual framework lies in prioritizing tangible advancements in patient health through the application of system-based approaches [98].

Within systems medicine, two overarching methodologies emerge: bottom-up and top-down approaches. The former focuses on developing intricate, quantitative mathematical models for specific subsystems. These models aim to elucidate the dynamic and nonlinear interactions among known components, enabling a deeper understanding and prediction of their behavior when subjected to perturbations [98]. On the contrary, top-down approaches utilize omics data to gain a comprehensive understanding of the components of a biological system. They seek to find molecular interaction networks and regulatory mechanisms from big data like genome-wide molecular data [99]. In pursuit of this end, prior knowledge on molecular interactions can also come into play.

A primary objective of top-down paradigm is the construction of system-wide, mainly static networks, such as gene co-expression networks and protein-protein interaction networks, offering a holistic perspective on functional and physical interactions within the biological system. The top-down approach proves invaluable for gaining a systematic understanding of diseases at the molecular level and, consequently, identifying potential treatments [99]. The more levels of biological information we include, the better systematic understanding of a disease on the molecular scale can be achieved. Integrating this multitude of biological knowledge is particularly a challenging task [100]. The computational methods used in the disease module identification and drug repurposing platforms presented in this dissertation follow the top-down systems medicine approach.

Network medicine, an offshoot of systems medicine, explores complex biological systems by employing principles drawn from graph theory. This term was coined by Albert-László Barabási in 2007 in his seminal work, "Network Medicine – From Obesity to the Diseasome" [101]. Barabási posits that biological systems share similarities with social and technological systems, characterized by intricate interconnected components governed by simple principles. These organizational principles can be thoroughly examined by representing systems as intricate networks, where nodes represent various biological elements such as genes, diseases, and phenotypes. The edges connecting these nodes depict the relationships between them, including physical interactions, similarities, shared metabolic pathways, shared genes, comorbidities, and more. Network-based approaches contribute to understanding human diseases and offer diverse applications, including the identification of disease genes and pathways, mechanistic insight into diseases, drug repurposing, drug target identification, personalized medicine, biomarker discovery, disease subtyping and stratification. These methodologies, along with the associated tools, collectively constitute the emerging field of network medicine [5,102].

2.4.1. Molecular and biomedical networks

In computer science, social and technological network research, and particularly in bioinformatics, graph representations of complex systems are widely employed [103,104]. The characteristics of graphs make them highly suitable for data integration applications, as they facilitate the storage and interconnection of data from diverse sources.

Our increasing knowledge of biological processes emphasize the intricate interplay among different components, including proteins, RNA, DNA, metabolites, and environmental factors, rather than relying on individual agents. Therefore, modeling cellular interactions or relationships as networks has emerged as a natural approach. Consequently, network biology has gained significant importance in bioinformatics research, focusing on the construction and analysis of networks that encompass diverse relationship types. These relationships, among others, include protein-protein interactions, transcriptional regulatory interactions, disease-gene associations, drug-target associations, and signaling pathways. In addition to molecular-based networks, networks based on biomedical knowledge, such as population-scale social and health interactions, can also shed light on disease biology and etiology. This type of data can be obtained, for example, from electronic health records, to indicate co-occurrences of diagnosed diseases (more accurately medical diagnostic codes) across patients [105].

The growing interest in developing graph visualization and analysis tools for biological networks underscores the benefit and power of using networks in computational biology. One of the widely used tools is Cytoscape [106], an open-source software platform for visualizing and analysis of molecular interaction networks and biological pathways. Cytoscape enables users to load pre-constructed interaction networks and perform straightforward analyses with its integrated functionalities, or delve into more advanced, specialized analyses using the apps developed for this versatile software platform. Certain Cytoscape apps are tailored to provide data querying functionalities by connecting to the existing widely used databases, such as DisGeNET [107], STRING [108], REACTOME [109], and Kyoto Encyclopedia of Genes and Genomes (KEGG) [110], allowing users to load and explore data in network formats.

In the following, I first present some graph terminology and then introduce important examples of molecular and biomedical networks underlying this dissertation.

2.4.1.1. Network modeling and terminology

As described earlier biological networks are modeled as graphs. A graph is a mathematical structure to describe pairwise relationships between objects. These

two terms are used interchangeably in this thesis. Graphs are defined in various ways depending on their application context (Fig. 2.8).

A graph is an ordered pair $G = (V, E)$ comprised of a set $V(G)$ of vertices or nodes and a set $E(G)$ of edges or links. Every edge is composed of a pair (u, v) of two endpoints in the set of vertices. If the graph is *directed*, the order of the pair indicates the direction (u being the source node and v being the target), while in an *undirected* graph the edges are not directed. Where the direction of the effect is unknown or nonexistent, undirected graphs are widely employed, e.g., to model protein-protein interaction networks. To represent relationships with an inherent directionality, such as metabolic and signaling pathways, directed graphs are used. A *directed acyclic graph* (DAG) is a graph with no directed cycles. A directed graph is a DAG if and only if it has a topological ordering (a linear ordering of vertices such that, for every directed edge (u, v) , vertex u comes before vertex v in the ordering). Due to their hierarchical structure, biological ontologies such as the Gene Ontology (GO) [111], and those disease terminology systems using ontologies (conceptual domain models) such as Mondo Disease Ontology (MONDO) [112] are modeled as DAGs.

A *bipartite graph* is a triplet $G = (U, V, E)$ where U and V are disjoint sets of vertices and E is a set of edges linking a vertex in U to a vertex in V . Bipartite graphs are employed for modeling networks where relationships map from one class of entities to another, such as disease-gene associations and drug-target interactions.

A *weighted graph* is a graph where each edge is given a numerical weight. It is common to model networks with pairwise similarities using weighted graphs, such as drug similarity, gene co-expression, and disease comorbidity networks. Edge weights may also be used to include confidence scores of interactions to take into account uncertainty in the detection or prediction method, e.g., gene-disease association (GDA) scores from DisGeNET database which are determined based on the level of evidence for associations and PPI scores from STRING database which serve as measures for confidence level of an interaction.

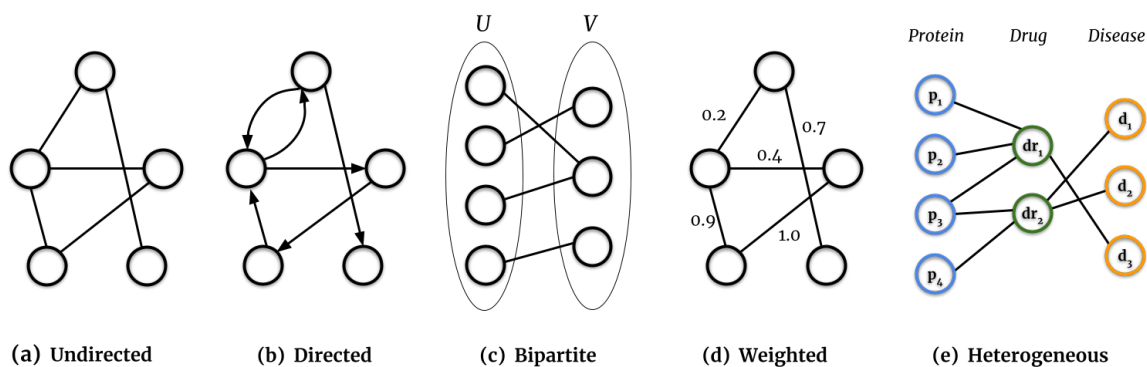


Figure 2.8 - Overview of different graph types.

A *heterogeneous graph* has nodes of different types, i.e., the set of nodes can be partitioned into disjoint sets $V = V_1 \cup V_2 \cup \dots \cup V_k$ where $V_i \cap V_j = \emptyset, \forall i \neq j$. Edges in heterogeneous graphs generally satisfy constraints according to the node types. One of the most common constraints is that certain edges only connect nodes of certain types. For example, in a heterogeneous graph of biomedical entities, there might be three different node types representing diseases, proteins, and drugs. Edges of type “has_target” are defined between drug nodes and protein nodes and can only occur between these two types of nodes. Similarly, edges representing “indicated_for” would only occur between nodes of type drug and those of type disease.

A *knowledge graph* is a structured representation of knowledge that utilizes a knowledge model consisting of interconnected descriptions of concepts, entities, relationships, and events while also encoding the semantics underlying the used terminology [113]. In other words, the entities in a knowledge graph are semantically enriched, which means they are associated with meanings and aligned with ontologies. This enables a computer to understand the context of an entity in the knowledge graph, its relationships to other entities, and its type (e.g., disease, gene, drug, or person) via the edges that connect the nodes. Also, in semantic graphs both vertices and edges can be attributed, i.e., they are annotated with attributes. A *metagraph*, or graph of types, can be defined as the grammar underlying a knowledge graph, i.e., representation of different relationship types between different entity types [114].

2.4.1.2. PPI networks

For over 20 years, information on interactions between proteins in humans has been collected and made accessible to the public through large databases. Databases like BioGRID [115], IntAct [116], HPRD [117], and STRING [118] contain hundreds of thousands of interactions across various species, curated from published research and derived from computational predictions. These databases allow for the creation of a network, known as an interactome, that includes all of these interactions as edges and proteins or genes as nodes. Since they lack direction and edge weight, PPI networks are typically represented as simple graphs. However, some protein interaction databases do include an interaction score [119]. STRING computes confidence scores for interactions based on the type of evidence contributing to the prediction. Such scores are not an indicator of the strength of the interaction but how likely STRING evaluates them to be true, given the available evidence. STRING considers interactions based on direct (physical) and indirect (functional) associations. Evidence types used include high-throughput lab experiments, automated text mining, genomic context prediction, amongst others [118].

Delivering an organized overview of which genes/proteins are interacting and/or functionally linked, interactomes have emerged as a key resource in bioinformatics. For instance, interactomes are used to find disease-associated mechanisms and biomarkers as well as to predict potential drug targets.

It is commonly claimed that biological networks, including PPI networks, resemble scale-free networks characterized by a power law degree distribution [120–122]. However, this assertion is a controversial one and some researchers argue that the scale-free nature of these networks is due to research bias or that other graph models are more fitting [123–126]. Initially, it was believed that high-degree proteins in PPI networks were crucial for cell survival [122,127]. However, this idea was later challenged by further studies. The affinity purification methods can also contribute to bias by favoring essential proteins that are more abundant [60]. There is also a correlation between node degree and the number of published papers about a protein, indicating that the importance of essential proteins as hubs may be partially due to their high level of research attention [50].

2.4.1.3. SARS-CoV-2 Virus-host interactome

Complementary to the human PPI network, maps of host-pathogen interacting proteins advances our knowledge about the molecular mechanism underlying viral and infectious diseases [65]. As described under the proteomics section, protein-protein interactions between SARS-CoV-2 and human proteins are identified using AP-MS technique. Integrating interactions between either host or virus proteins and other host proteins into one network, such as SARS-CoV-2 virus-host interactome integrated in the CoVex web tool (my first publication), enables researchers to investigate downstream host proteins that could play important role in the viral replication cycle and could be potential drug targets [11]. Other efforts like IMEx-wide initiative [128], compile molecular interaction data related to SARS-CoV-2 and other viruses in the Coronaviridae family, along with human protein interactions that may be pertinent to the etiology of the disease. The IMEx coronavirus interactome is available as part of IntAct molecular interaction database and has been evolving as more studies on interactions evidence have been published since the outbreak of the COVID-19 pandemic [129]. This ongoing work also incorporates novel interactions and details of known interactions such as the effects of variants.

2.4.1.4. Diseasomes

Earlier studies aimed at compiling disease-gene associations primarily focused on individual diseases, exploring the relationships among genes implicated in specific disorders [130]. In this context, Goh et al. [91] tried to improve the single gene-single disorder traditional approach by developing a conceptual framework that systematically links all genetic disorders with the full list of disease genes.

Their effort led to the construction of the so-called “diseasome” network, which gives a global view on the combined set of all known diseases and their associated genes [91]. More formally, diseasome is a bipartite graph that comprises two distinct sets of nodes, one of gene node type and one of disease node type. The disease set represents all identified genetic disorders, while the other set represents all known disease genes within the human genome. Connections or links between a disorder and a gene are established when mutations in that gene are associated with the corresponding disorder [91]. From diseasome, a disease-disease network projection is inferred, where nodes represent diseases, and there is an edge between two diseases if they have at least one associated gene in common. In this dissertation, we refer to this disease-disease network as diseasome.

The initial diseasome used the data from OMIM database [22]. The gene-based diseasome generated for the last publication presented in this dissertation [131] builds upon Goh et al.’s work [91], incorporating additional disease-gene data from other resources. In some other studies, the idea of global view on disease relationships was also applied to other types of disease-associated data such as pathways. Li et al. [92] studied connections between diseases and built a pathway-based diseasome by associating diseases to biological pathways through disease genes. In my last publication, where we investigated the bias in some network medicine studies which is the result of using inadequate disease definitions in current medicine discipline, we additionally generated various types of diseasomes based on a variety of disease association data types, such as disease-symptom, disease-variant, drug-indication, and comorbidity relationships between diseases.

Disease-gene associations

OMIM provides a catalog of human genes, genetic disorders and traits and was the only source of disease-gene associations used for the construction of the original diseasome introduced by Goh et al. [91]. OMIM particularly focuses on the molecular relationship between genetic variation and phenotypic expression and is based on manually curated data. The gene-based diseasome presented in my last publication [131] is an expanded version of the original diseasome, enriched with disease-gene data sourced from DisGeNET [107]. This comprehensive database aggregates disease-gene associations from a multitude of databases, including UniProt [132], CTD [133], Orphanet, ClinGen [134], Genomics England [135], CGI [136], and PsyGeNET [137]. Unlike OMIM, DisGeNET offers a broader spectrum of disease-gene associations, encompassing genetic variation, causal mutations (mutations known to cause the disease), modifying mutations (mutations known to modify the clinical presentation of the disease), statistical associations (without evidence of causality), and chromosomal rearrangement [138].

2.4.1.5. Comorbiditome

Although significant advancements have been made in molecular profiling and high-throughput omics technologies, many available resources fail to consider the vast and regularly updated phenotypic information we have for humans, specifically in the form of patient clinical histories [139]. Notably, hospitals and insurance companies gather comprehensive records for millions of patients, which include details on disease associations and progression as well as prescribed medications. These population-based data hold valuable information that, when combined with molecular and genetic data, can aid in unraveling the molecular causes of diseases [140]. Some countries like the United Kingdom and Denmark have central and long-established electronic health record (EHR) systems [141,142], while others like Germany are still at the beginning of the road [143]. Due to lack of extensive medical records accessibility, there are not many population-based disease association databases, from which comorbidity associations among diseases could be inferred [139]. A comorbidity relationship arises when two diseases co-occur to the same individual significantly to a greater extent than what would be expected by chance. Disease comorbidity networks, so-called “comorbiditomes”, can be constructed based on the disease diagnoses data. Different studies defined different metrics to quantify comorbidity relationships [144], some even analyzed temporal comorbidities and created disease trajectories based on longitudinal patient data [145,146].

2.4.1.6. Drugome

As described previously, some computational drug repurposing approaches utilize the knowledge from the relationship between drugs. This connection between drugs can be, for example, inferred from shared indication, shared target protein, and similarity based on chemical structure. As a result, global views on drug relationships can be achieved by constructing different drug-drug networks based on the aforementioned shared data. We refer to such drug-drug networks as “drugomes”.

2.4.2. De novo vs. traditional enrichment methods

Integrating prior knowledge can be advantageous for the computational analysis of omics data in systems medicine approaches. Over the past decades, extensive research in molecular biology has yielded a wealth of knowledge on molecular interactions, functions, and pathways. Leveraging this knowledge can significantly improve computational methods and particularly enhance interpretability. The databases providing this prior knowledge are often publicly accessible which facilitates the integration of this type of data to the computational analyses. They include but are not limited to the pathway databases

like KEGG and Gene Ontology (GO) [147], as well as protein-protein interaction databases, such as BioGRID, IntAct, STRING, and the Integrated Interactions Database (IID) [148]. In the following, two general types of analysis methods using prior knowledge are described.

Gene set and pathway enrichment analysis methods

These methods aim to find known pathways that are significantly dysregulated between two clinical conditions. To this end, gene set over-representation analysis (GSORA) methods and gene set enrichment analysis (GSEA) methods are applied [149]. The former is a statistical method determining if genes from a pre-defined set (e.g., those belonging to a specific GO term or KEGG pathway) are over-represented, i.e., present more than expected by chance, in a set of under or over expressed genes. GSEA is one of the functional scoring methods where the entire set of gene-level statistics is used to identify enriched gene sets. GSEA employs a statistic similar to Kolmogorov-Smirnov test to quantify the degree to which genes in a gene set are overrepresented at the extremes of the entire ranked list, indicating over- or under-expression [150].

De novo network enrichment methods

Disease module identification methods (DMI), also known as de novo network enrichment (DNE) or active module identification methods, aim to detect a disease module - a connected subnetwork within the human interactome that associates with a particular disease and either exhibits statistically significant enrichment or fulfills specific connectivity criteria [16]. The disease module concept arises from the observations established from different studies that disease genes are not distributed randomly, but rather exhibit a tendency to be closely interconnected or reside in close proximity within the interactome [91,151].

Unlike traditional enrichment analysis, as previously discussed, DMI adopts a more data-driven approach to extract condition-specific subnetworks. The traditional enrichment methods are dependent on pre-defined and curated pathways or gene sets [152], limiting their ability to uncover novel disease mechanisms. DMI methods, on the other hand, first build “active” subnetworks through the projection of experimental data, mainly transcriptomic or genomic profiles, onto a global molecular interaction network like PPI networks. Subsequently, an objective function is applied to evaluate candidate subnetworks, employing efficient heuristics to find locally optimal solutions [16]. In their study, Batra et al. [152] compared various state-of-the-art DNE methods and reached the conclusion that the most suitable strategy for identifying optimal subnetworks depends on the specific application at hand.

In our review paper “Network-based approaches for modeling disease regulation and progression” [16], we presented an overview of recent network-based methodologies and their main concepts aiming to identify disease modules or

potential mechanisms, allowing for a deeper understanding of diseases which potentially leads to better drug discovery and precision medicine. In the following, I provide a summary of different types of the DNE methods discussed in this review paper. We can classify DMI methods roughly into four categories:

Aggregate score methods (ASM) - These methods first assign some scores to each gene, typically derived from a case-control experiment, such as differential expression analyses, and can be a test statistic, P value or fold-change. ASM methods aim to find a connected subnetwork with maximum aggregated score.

This type of analysis was first introduced by Ideker et al. in 2002 [153], where a simulated annealing algorithm was utilized to identify a subnetwork with the maximum aggregated z -score. In some studies, scores are assigned to edges instead, e.g., based on co-expression [154]. In this dissertation, I developed a disease module identifying algorithm based on the Steiner tree (ST) concept. ST problem seeks to find a tree of minimum cost connecting a given set of terminal nodes [155]. STs can be considered as extensions of shortest paths in scenarios involving more than two endpoints. In the context of network enrichment these points correspond to the disease genes known a priori.

Score propagation methods (SPM) - Similar to ASM, these methods also first assign a score (or “heat”) to each node in the network and then simulate the propagation of the score throughout the network over time, leading to the accumulation of the signal in signal-rich subnetworks. The majority of SPM methods are based on either heat diffusion [156,157], random walks [158,159], or network expansion starting from seed genes [160,161]. In contrast to ASM methods, SPM does not extract a connected subnetwork. Instead, by simulating network flow, it reassigns priority to genes using the network information such as network topology which is beyond just connectivity.

Module cover (MC) - These methods follow a two-step process: Firstly, a set of key genes is chosen, usually genes that exhibit significant differential expression or known disease genes. Next, a connected subnetwork with a high density of key genes is extracted, typically with certain constraints such as limiting the total number of genes or the inclusion of non-key genes [162].

Some methods, like KeyPathwayMiner [163], evaluate differential expression on a per-patient basis and permit exceptions at both the gene and patient levels. By dissociating the selection of key genes from subnetwork extraction, these methods avoid presumptions about the underlying data, enabling the choice of an appropriate statistic depending on the dataset. However, the selection of significance cutoff for genes, the desired module size or number of exceptions, has a substantial impact on the results, making the application and interpretation of MC methods more challenging [152].

Machine learning-based approaches - These approaches mainly employ clustering methods, an unsupervised approach, to identify clusters of differentially expressed genes that exhibit strong connectivity and co-expression. Methods can adopt one of two approaches: traditional network clustering, where the edge weight signifies the similarity among genes in the molecular profile [164], or network clustering strategies that directly cluster differentially active genes in the network [165].

The majority of current DMI methods heavily depend on case-versus-control annotations. However, unsupervised approaches offer the advantage of not only identifying disease modules but also simultaneously clustering patients into subgroups, providing additional insights [16]. Biclustering Constrained by Networks (BiCoN) is an unsupervised approach based on the Ant colony optimization algorithm [44]. The approach utilizes a heterogeneous network comprising patients and genes. In this network, genes are connected to patients through expression data, while they are linked to each other through PPIs.

Network pharmacology and disease modules

The current “one disease-one target-one drug” dogma in drug discovery impedes effective treatments for complex diseases [166]. Due to intricate molecular and environmental interactions, many diseases are influenced by complex factors. Consequently, merely treating a single component, target, or pathway may not be sufficient to disrupt the underlying mechanisms responsible for the disease [167]. Network pharmacology as a new therapeutic branch of systems and networks medicine considers that mechanisms underlying complex diseases involve a subnetwork of interconnected genes rather than just one gene or protein [168]. Targeting this subnetwork, so-called disease module, instead of only one protein has shown the potential to be a more effective therapeutic intervention [169]. This can be achieved by either one drug targeting multiple proteins in the module or multiple drugs targeting different proteins and acting synergistically. This method can also lead to reduced individual drug dosage and possible side effects due to mechanism-based synergy [166]. Therefore, drug repurposing approaches which include steps in their pipeline to identify drugs targeting multiple proteins within disease modules rather than single protein are promising to achieve more effective and precise treatments.

2.4.3. Disease ontologies and vocabularies

A disease ontology is a formal representation of knowledge about diseases and disorders in a structured and hierarchical fashion, i.e., higher level terms have more general meanings than their lower level counterparts. It typically includes a set of concepts, objects, and other entities that represent various aspects of diseases, such as their etiology, pathogenesis, symptoms, and treatments, as well

as the relationships and dependencies among them. The ontology may also categorize diseases based on their characteristics, such as their anatomical locations, genetic basis, or clinical manifestations. Different disease ontologies may have different design principles, coverage, and levels of granularity, depending on their intended use cases and the expertise of their developers. Some examples of disease ontologies include Human Disease Ontology (DO) [170,171], MONDO [112], and Orphanet Rare Disease Ontology (<http://www.orpha.net>). The term “disease vocabulary” used in this dissertation encompasses disease ontology as well as other disease terminology systems which do not have an ontology-based structure, such as International Classification of Diseases (ICD) codes [172], Medical Subject Headings (MeSH) terms [173], Unified Medical Language System Concept Unique Identifiers (UMLS CUIs) [174], Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) [175], and OMIM [176]. Although in the design of some disease vocabularies interoperability for the purpose of data integration was taken into consideration, among the numerous existing disease vocabularies, there is not a conclusive standard for encoding diseases while addressing requirements of information exchange. Each of these vocabularies is designed for a particular purpose, hence, they only partially overlap and often disagree in the mapping approach. This makes it difficult to align them with each other and/or with other knowledge sources.

Endotypes

The foundation of how diseases are currently defined, independent of the disease ontology of choice, is mainly based on symptoms (phenotypes) and/or involved organs. Examples of symptom-based disease definitions: hypertension, defined by elevation in the blood pressure [177] or hyperlipidemia, defined by abnormally elevated levels of any lipids in blood. Examples of organ-based disease definitions: Kidney failure, defined as an acute or chronic condition that is characterized by the incapability of the kidneys to effectively filter the blood.

One of the goals of the network medicine field is to uncover pathomechanisms driving diseases and consequently replace the currently ill-defined disease classifications and definitions by a mechanistically grounded disease vocabulary [168,178,179]. The distinct molecular mechanisms underlying the disease phenotypes are called *endotypes* [168,180,181]. Replacing phenotype-based disease definitions by endotypes is a significant step towards disease-modifying treatments rather than symptom-alleviating treatments. This newly proposed endotype-based disease vocabulary does not require redefining semantic relationships between existing disease terms, in other words, the aim is not to build yet another disease ontology, but discovering currently unknown molecular disease mechanisms and breaking down fuzzy and umbrella disease terms such as coronary artery disease into endotypes which are characterized at a molecular level [131].

DMI methods explained earlier in this chapter aim to uncover previously uncharacterized molecular mechanisms which potentially assist in defining diseases based on endotypes.

2.4.4. Bird's-eye-view vs. close-up network medicine

At one end of the spectrum of network medicine approaches, we have *close-up* methods employing molecular data for well-characterized patient cohorts to study a specific disease that can lead to discovering novel mechanistic insights. At the other end of the spectrum, we have bird's-eye-view (BEV) network medicine approaches using large-scale disease association data combined from several data sources to infer knowledge about diseases by investigating their relationship at network levels [131]. In spite of the promising and translational findings of close-up approaches, many studies showed biases in the data used by BEV approaches. These studied biases are mainly related to proteins and genes in the context of networks (as one end of an edge or association in a network) and less attention has been paid to biases that are caused by diseases (as another end of an association). As mentioned earlier, network medicine aims at finding underlying disease mechanisms and replacing disease definitions by their mechanisms, at the same time, it uses ill-defined disease definitions to achieve this goal. We postulated that using the data which is based on fuzzy disease definitions can introduce bias and formulated two testable hypotheses, in global- and local-scale, that I investigated in the fourth publication. I tested the hypotheses by quantifying the pairwise similarity of several diseasesomes and drugomes constructed from different types of association data (details can be found in the General Methods chapter).

2.5. Data integration for in silico and network-based drug repurposing

2.5.1. Challenges

Data integration in life sciences can be particularly challenging due to the increasingly large and complex nature of data sets, commonly referred to as "Big Data". These data sets are often too massive to be processed by a single machine and require specialized tools and distributed computing resources. Moreover, the data is spread across various databases, each following its own conventions, using its own vocabularies and is available in different data formats, which adds to the complexity of integrating and analyzing the data. In the following, I present a non-comprehensive overview of the challenges with data integration.

Data heterogeneity, semantics and syntax

As the level of data integration increases, the main challenge is the heterogeneity of data, which refers to the differences in data types. The more diverse the data types are, the higher the chances of encountering mismatches when trying to integrate them. When mixing data from various sources, issues arise not only with the format or syntax of the data but also with its underlying meaning or semantics [182].

Inference of equivalence and mapping

Coalescing data from different sources entails identifying the various data ‘types’, e.g., drugs, genes, or diseases, and establishing equivalence or mappings between corresponding entries in each source. This can be a relatively easy task if there are standardized accessions and unique simple keys, such as HUGO Gene Nomenclature Committee (HGNC) symbols and entrez IDs for genes, but it is more difficult for complex entities like diseases, which often lack such clear and standardized identifiers [182]. The need for integrating disease-related information from different resources has resulted in a large number of mapping systems between different disease vocabularies. As elaborated in the disease ontology section, these mappings lack completeness, accuracy, and precision.

Integrating cases for which mapping from one vocabulary to another is one-to-one is a straightforward process. For instance, consider achondroplasia, the most prevalent form of chondrodysplasia. In the Orphanet Rare Disease Ontology (ORDO), it is assigned the disease ID Orphanet_15, while in the MeSH hierarchy, it has the disease ID D000130 and in the MONDO system has the ID 0007037. All three of these entries refer to the same disease within different terminology systems. In a graph-based data integration exercise, this disease could be represented as a single node encompassing all three labels, Orphanet_15, D000130, and 0007037. In some instances, a single disease ID from one vocabulary may be mapped to multiple disease IDs in another vocabulary. This non-unique one-to-many (1-to-n) mapping complicates the integration process, necessitating the selection of a single, definitive vocabulary system to map all diverse disease vocabularies to. For example, OMIM defines Choroid plexus papilloma with ID 260500 as choroid plexus tumors with neuroectodermal origin, ranging from benign choroid plexus papillomas to malignant choroid carcinomas. However, there are two distinct disease IDs for the benign and malignant form in ICD-10 (D33.1 and C71.7, respectively) and MONDO vocabularies (0009837 and 0016718, respectively).

Updating and Metadata management

The datasets or databases are subject to frequent updates, and the changes between each release can be significant, either in terms of the data format or the data content. In case of data format or data structure change, before updating the integrated database, these changes need to be captured to adapt parsers and methods previously used for the integration. In order to effectively handle

differences in data sources during database integration, it is essential for the integrated data source to include metadata, which refers to information that describes the data, such as data provenance (i.e., origin and timestamp of data). In projects where integrated datasets get continuously updated, to be able to reproduce the results derived from the database, or compare results from different versions of a database, the need for provenance becomes even more critical.

Data access and representation

When it comes to data integration, it is important to think beforehand about the method of access, and the way in which the data will be queried or visualized, as this may require different types of data representation. There are various ways to access biological databases, such as through Representational State Transfer (REST) services, SQL databases, flat files via File Transfer Protocol (FTP), and many others. It is often essential to have multiple access points to fulfill different purposes [182].

After identifying and evaluating the challenges associated with the data integration as one of the initial steps in the drug repurposing project, it is crucial to choose a data integration platform that can help achieve all the project's objectives.

2.5.2. Database models

Amongst many database models used for data storage and data mining in computational tasks, there are three models used more commonly in computational biology: in-memory, relational and schema-less databases [183]. When exploited to their full potential, in-memory formats, like those used by Cytoscape, Ondex [184], and Gephi [185], are quick. However, they are optimized for analyzing small data sets that can fit in a single machine's memory. Therefore, in-memory approaches are constrained by the availability of memory rendering them infeasible to store graphs which are the preferred representation of integrated data sets in bioinformatics. Graphs frequently comprise a vast number of nodes and edges, making it impossible to represent them in RAM, resulting in their storage in databases [186].

Most relational databases use the structured query language (SQL) for managing the data. Traditionally, MySQL and PostgreSQL have been the go-to options for storing data. These relational databases operate by organizing data into tables, which are composed of rows and columns. Each row can be viewed as an object that possesses specific attributes or properties, represented by the columns [187]. With a history dating back to the late 1960s, relational databases have been heavily researched and optimized for efficient querying [186]. While the strengths and weaknesses of relational databases are well understood, their shortcoming in capturing necessary semantics constrains their applicability. Schema-based data

models impose constraints on how data can be stored, necessitating manual redesigns to accommodate new relationships. Relational databases excel in handling complex queries and set operations aggregating data. In contrast, graph databases are optimized for highly interconnected data [186].

The term 'NoSQL' is used to describe databases that lack a fixed schema (schema-less), including key/value stores (e.g., Apache Cassandra www.cassandra.apache.org), document stores (e.g., MongoDB www.mongodb.com), and graph databases (e.g., Neo4j www.neo4j.com). These databases are gaining popularity due to their scalability and flexibility, in contrast to the more traditional relational approach. In graph databases, edges are represented as directed pointers between nodes that can be traversed in constant time, depending on the implementation. When analyzing data interconnectivity or topology, graph databases become especially relevant, and they are optimized for graph traversals, such as algorithms for finding the shortest path [186].

Most graph databases do not possess a declarative query language. Neo4j is an exception with its query language, Cypher [188], inspired by SQL and allowing users to construct expressive and efficient graph queries. One of the distinctive features of Cypher is its ability to match patterns and relationships using a visual approach. This approach was inspired by an ASCII-art type of syntax that involves using rounded brackets to denote circular (nodes) and `-[:ARROWS]->` to represent relationships in the form of `(nodes)-[:ARE_CONNECTED_TO]->(otherNodes)`.

Knowledge graphs bring together elements from multiple data management approaches. They incorporate aspects of databases, allowing for exploration through structured queries. Additionally, they exhibit characteristics of graphs, enabling analysis similar to other network data structures. Moreover, knowledge graphs possess qualities of knowledge bases, as they incorporate formal semantics that facilitate data interpretation and the inference of new information [189]. A knowledge graph can be modeled with different database models, such as Neo4j, MongoDB, graphml format and SQL. The database models and data sources used for constructing the knowledge graphs underlying this dissertation are introduced in the General Methods chapter.

3. General Methods

This chapter gives a brief summary of the methodologies applied in the publications forming this dissertation.

The overview includes: 3.1) different components of drug repurposing platforms; 3.2) data integration and network construction from public databases and the Estonian biobank; 3.3) network medicine algorithms used in the CoVex and NeDRex platforms that were adapted or de novo developed (like MuST) for the purpose of disease module identification and drug ranking; as well as 3.4) the network metrics employed for the similarity analyses of diseasesomes and drugomes (the fourth publication) to test the bias introduced to network medicine studies by inadequate disease definitions. The complete description and details can be found in the Methods section and Supplementary Information of the publications [11,12,131] (Appendices A.1, A.2, and A.4).

3.1. Overview of the drug repurposing platforms

The network medicine platform, NeDRex, presented in the second publication [12] has four main components including NeDRexDB, NeDRexApp, NeDRexAPI, and network medicine algorithms. NeDRexDB, integrates data from various biomedical databases relevant to the task of drug repurposing into a unified knowledge graph. To make our disease module identification and drug repurposing methods more accessible for biomedical researchers, we opted for implementing the user interface part of the platform as a Cytoscape 3 App, called NeDRexApp.

Cytoscape 3 is built on the Open Service Gateway Initiative (OSGi) framework, and its inherent modularity, enforced by OSGi, enhances extensibility. This quality aligns well with the design goals of the NeDRex platform, where a modular structure facilitates incorporating new algorithms and exploration functions. NeDRexApp provides implementations of network-based algorithms in the back-end via NeDRexAPI. The app functions as a user-friendly front-end interface, providing access to NeDRexDB. It enables users to build customized heterogeneous networks, execute network medicine algorithms, and visually explore the resulting data. The NeDRexDB knowledge base can be accessed via Neo4j endpoint as well (Fig. 3.1).

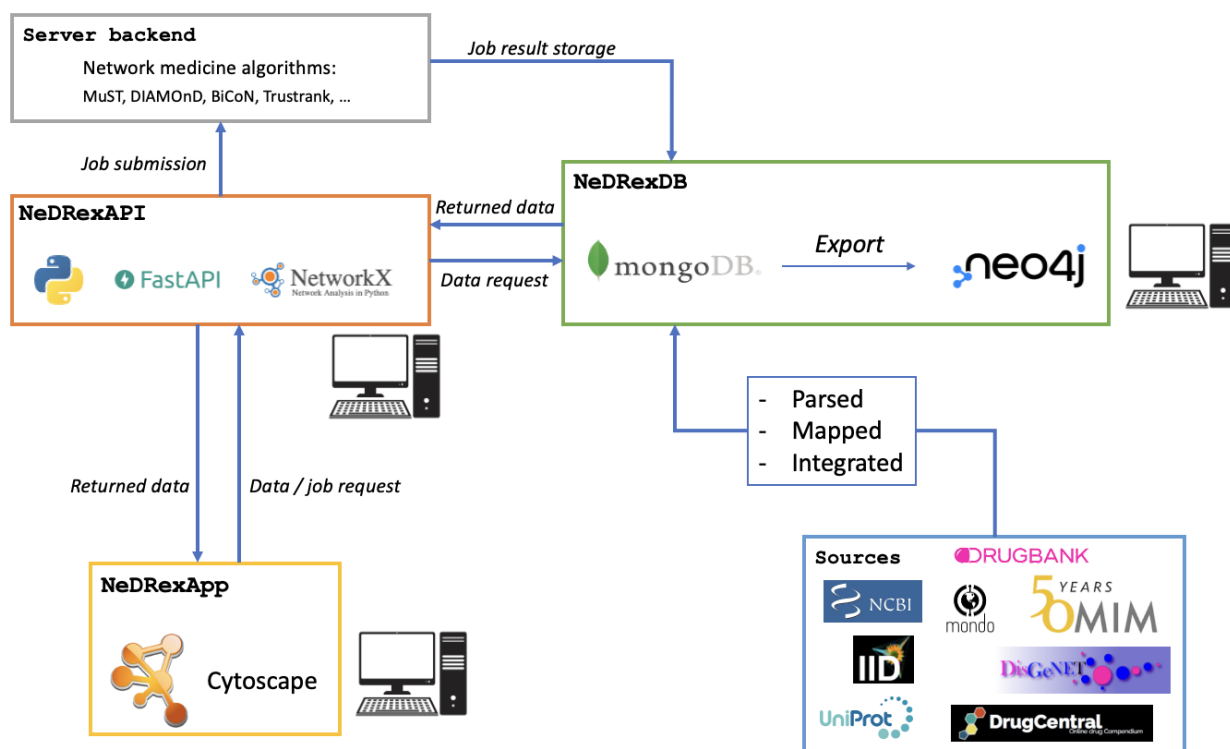


Figure 3.1 - The NeDRex framework, its main components and the connections between them.

The CoVex web platform, presented in the first publication, is composed of four main components: the SQL database, the backend to build the web API, the network medicine algorithms, and the frontend for network visualization. The implementation details can be found in the original publication [11] (Appendix A.1).

3.2. Data integration and network construction

One of the primary steps in developing network-based drug repurposing platforms of CoVex and NeDRex was to integrate relevant databases into a knowledge graph. For CoVex, we utilized a relational database implemented in PostgreSQL. NeDRexDB is a knowledge graph modeled as Neo4j and MongoDB with the possibility to export to any graph format such as graphml. There were two main factors behind the selection of MongoDB as the database model. Firstly, MongoDB's flexible schema allows for easy addition of new attributes to database documents while also providing the option to selectively enforce specific guarantees [190]. Secondly, MongoDB offers a wide range of operations for querying and updating data, thus greatly supporting data integration efforts [191].

A comprehensive list of data sources and the types of integrated data for the CoVex and NeDRex platforms [11,12] can be found in the respective publications (Appendices A.1 and A.2). For the fourth publication where we evaluated the bias introduced by disease definitions in network medicine studies, we created a more

complete version of the disease-disease (diseasome) network initially introduced by Goh et al. [91], by including disease-gene data from additional data sources. In a similar fashion, we defined and generated other types of diseasomes based on different types of disease association data, such as disease-symptom (also referred to as phenotype), disease-variant, drug-indication, and comorbidity relationships between diseases. Inspired by the notion of the diseasome utilizing bipartite graphs of drug-target and drug-indication, we constructed drug-drug (drugome) networks. The construction of different types of diseasomes and drugomes is illustrated in Fig. 3.2.

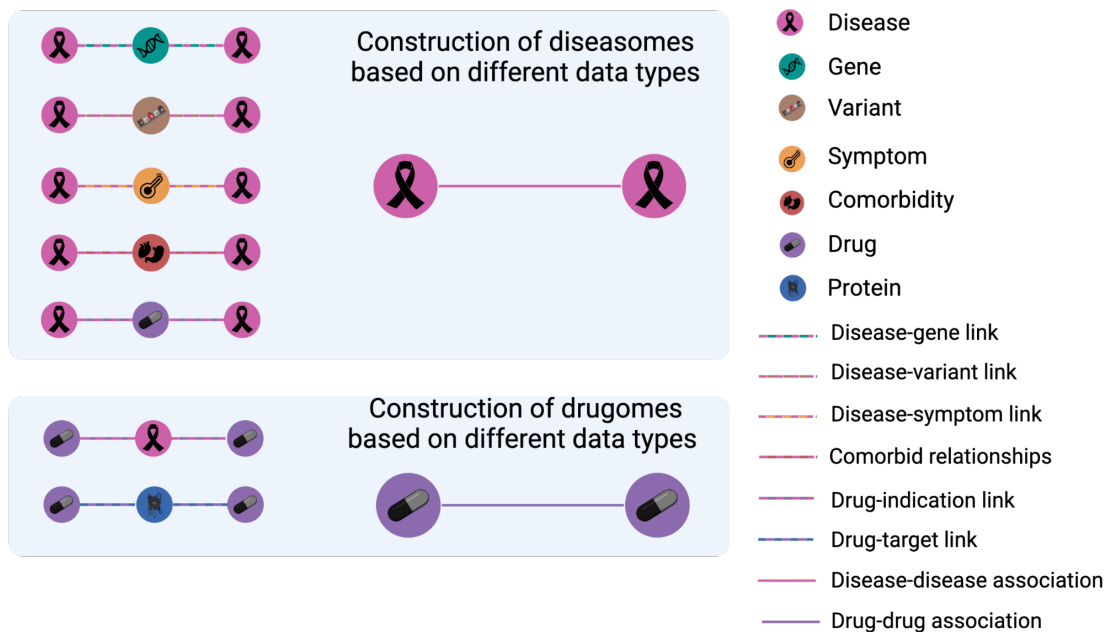


Figure 3.2 - Construction of diseasomes and drugomes based on different data types. Figure adapted from Sadegh et al. [131]. Permission granted by the authors.

Most of the networks were constructed based on the data from publicly available databases. Only for the comorbiditome, the disease-disease network based on comorbidity data, we used the ICD-10 diagnoses stored in the health records available from the Estonian population-based biobank.

The full metagraph of the NeDRRexDB knowledge base together with additional data used for the analyses in the fourth publication is illustrated in Fig. 3.3.

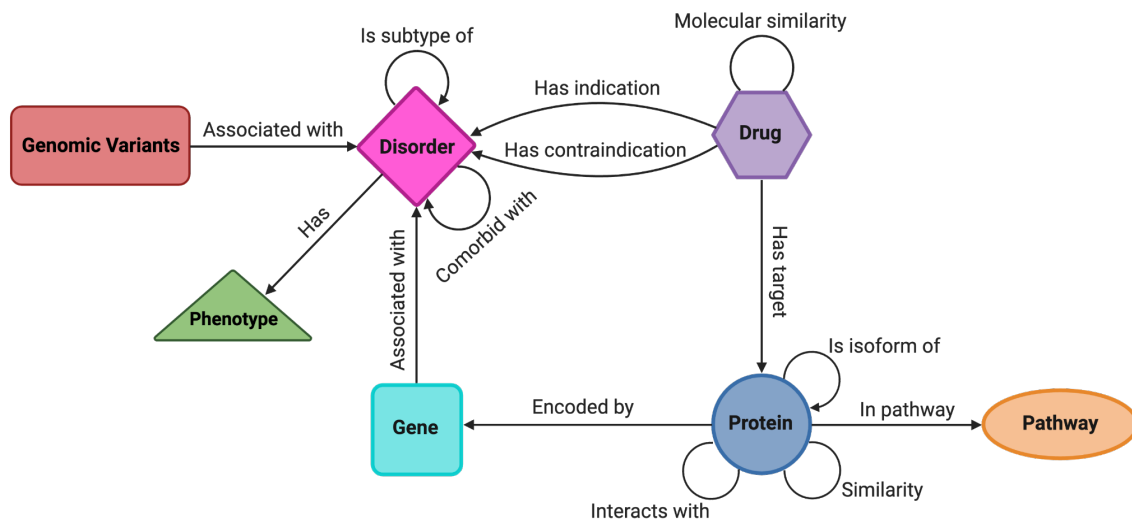


Figure 3.3 - The NeDRexDB metagraph appended with data types used for network similarity analyses in the fourth publication. Figure modified from Sadegh et al. [12] and created with BioRender.com. Permission granted by the authors.

A significant challenge we faced during the data integration process was the discordant use of various disease vocabularies (identifiers) across databases containing disease association data. For instance, both DrugCentral [192] and CTD databases contain drug indication data, but while DrugCentral uses SNOMED CT to denote the indications (diseases), CTD uses MeSH terms. In the context of network medicine applications and to investigate underlying mechanisms of diseases, we need to map data to a common target vocabulary if we want to jointly leverage the disease related data from various data sources. This often leads to information loss due to unmappable terms [131]. MONDO is the result of a big community of experts working together to unify multiple disease resources by providing a logic-based structure [112]. Based on our assessment, it showed the highest mappability to other disease vocabularies among the available options [131]. Consequently, it was chosen as the target vocabulary for the integration of the databases within the NeDRex platform. In the fourth publication, we also assessed the impact of annotating the data with disease vocabularies of varying levels of granularity on the results. We conducted the analyses using MONDO IDs and UMLS CUIs for finer granularity, and ICD-10 three-character codes for coarser granularity, as node IDs in the generated networks. This task required the construction of the networks in different disease vocabulary systems [131].

3.3. Network medicine algorithms

For the first publication, CoVex, we have adapted several already established algorithms and integrated them into the web tool, each based on distinct paradigms. This selection of algorithms aims to offer targeted exploration options

for diverse research inquiries and hypotheses related to the COVID-19 disease and therapeutic drugs.

These algorithms include: modified versions of closeness and betweenness node centrality measures [193], network proximity [87] based on the distance between drug targets and disease related genes, KeyPathwayMiner [194] a de novo network enrichment method to identify condition-specific key pathways, and TrustRank [195] a variation of Google's PageRank algorithm that involves the iterative propagation of "trust" from seed nodes to adjacent nodes by utilizing the underlying network structure. We also developed a new algorithm based on the Steiner tree concept, called MuST, to identify drug targets for COVID-19, described more in detail later in this section. All the algorithms require hypothesis-driven starting points, so-called seeds, that can be either viral proteins, human proteins or drugs. Based on the initial hypothesis and selected seeds, integrated systems medicine algorithms find connecting paths from viral proteins to drugs using host proteins from the human PPI interactome as proxies. The algorithms are grouped into two categories of drug target and drug candidate discovery.

Regardless of the network analysis method used, high-degree nodes, i.e., hub proteins having a large number of interactions in the PPI network, appear in the obtained results with a higher likelihood. Since hub proteins are also more likely to be part of multitude mechanisms and are not particular to the mechanisms of the studied disease, we introduced *hub penalty* to the algorithms to mitigate this bias. By this means, we penalize high-degree nodes by incorporating the degree of neighboring nodes as edge weights in the optimization step.

For the second publication, NeDRex, we built on the methodology of CoVex and integrated disease module identifying methods benefitting from the integration of prior knowledge, like PPI networks, and overlay molecular profiles on these networks to derive new candidate disease mechanisms. We integrated BiCoN [44] which is an unsupervised method and performs simultaneous patient clustering and protein module extraction to obtain a potential mechanistic explanation of a studied condition. We integrated DIAMOND [196], an established method designed based on the systematic analysis of connectivity patterns of disease-associated proteins within the human interactome. Additionally, we improved the previous implementation of the MuST algorithm in CoVex for disease module identification in the NeDRex platform to obtain more robust results.

In the NeDRex workflow, the next step after identifying disease related modules, is to prioritize drugs targeting proteins in the derived modules or in the vicinity of them. Two algorithms from our experience with CoVex showed to be more beneficial for drug ranking, namely TrustRank and Closeness centrality. These algorithms need initial seeds as input. These can be all proteins returned as result in the previous step (disease module identifying) or a selection of them reliant on the expert knowledge of users. This is where the human-in-the-loop concept can

come into play, enhancing predictions. Many effective drugs function by targeting multiple proteins rather than a single target. Both implemented drug ranking algorithms in NeDRex incorporate this principle. Drugs having more connections to the disease module proteins, i.e., targeting a higher number of proteins in the disease module, are scored higher. This strategy aligns with the concept of network pharmacology, discussed in the Background chapter, suggesting that therapeutic interventions can be more effective when targeting a subnetwork of proteins (disease module) rather than individual proteins.

A detailed description of all the network medicine algorithms implemented for CoVex and NeDRex platforms can be found in the original publications [11,12] (Appendices A.1 and A.2).

Multi-Steiner Tree (MuST) algorithm

The Steiner tree (ST) problem is a combinatorial optimization problem seeking a tree of minimum cost connecting a given set of terminal nodes. This problem in graphs can be viewed as an extension of two more well-known combinatorial optimization problems: the (non-negative) shortest path problem and the minimum spanning tree problem.

There is a tendency for functionally related genes to exhibit proximity within the PPI networks. Furthermore, it has been observed that the distribution of pairwise shortest paths among known disease genes has a significant leftward shift compared to what would be expected randomly [151]. Therefore, a rational hypothesis suggests that the shortest paths connecting these disease genes coincide with underlying molecular pathways of diseases [102].

Considering that STs can be seen as extensions of shortest paths in scenarios involving more than two terminals (equivalent to seeds in our methodology), it is reasonable to anticipate that a disease module constructed using STs would encompass a substantial portion of the molecular pathways relevant to the disease [15]. Exactly solving the ST problem is NP-hard, but several efficient approximation algorithms exist, such as the 2-approximation by Kou et al. [197].

Since PPI networks used in our settings are rather dense (have high edge-to-vertex ratio), solutions to the Steiner tree problem are usually non-unique. Therefore, we return the final module as the union of multiple solutions to ST problem, hence the name 'Multi-Steiner tree'. I also co-authored another publication, where we further improved the implementation of MuST presented in this dissertation to be more robust by enumerating maximally diverse prize-collecting Steiner trees [15].

Statistical validation

In order to evaluate the statistical significance of both types of the results, i.e., identified disease modules and predicted drugs, generated by various algorithms

within NeDRex, we have incorporated three validation strategies that rely on empirical P -values. A list of reference drugs, that is the drugs indicated for the disease under study and/or drugs undergoing clinical trials for the treatment of disease, is needed for all three validation methods. The quality of this list with regards to completeness and false positives affects the P -value results.

The first method is the validation of predicted final drugs without taking into account the previous disease module identification step. We generate a high number of random ranked lists of drugs with the matched list's size to that of the reference drug list. We then compute the discounted cumulative gain (DCG) [198], reflecting to what extent a ranked list is in accordance with a reference list, for the predicted list of drugs and all random counterpart lists. The empirical P -value is then computed by counting the number of random drug lists whose DCGs exceed the DCG of the drug list predicted by NeDRex divided by the total number of considered random lists. The second method is designed to validate the predicted disease module in terms of druggability. For this method, we generate a high number of random mock disease modules of matching size and topology to the disease module identified by NeDRex. The precision of a module is defined by the number of reference drugs targeting the module divided by the overall number of drugs targeting the module. Similar to the first method, the empirical P -value is computed by counting the number of mock modules with higher precision values than the predicted disease module divided by the total number of simulated random mock modules. The third method is a joint validation method and takes into account both steps of the NeDRex workflow. This method is computationally similar to the second method with the difference in computation of precision of the predicted result. Here, we calculate the precision by counting the overlap between the reference drugs and the predicted drug list divided by the overall number of drugs in the list. A more formal description of the validation method can be found in the Methods section of the publication [12] (Appendix A.2).

3.4. Testing the bias introduced to network medicine studies due to inadequate disease definitions

The present large-scale disease databases use phenotype- or organ-based definitions for diseases. At the same time network medicine seeks to improve these ill-defined classifications and find underlying mechanisms for diseases to be the new definition system for diseases. However, many network medicine studies which are based on BEV approaches (defined in the Background chapter) and use such large-scale disease data runs the risk of replicating the biases inherent in these inadequate disease definitions. Therefore, these approaches assume that the biases originating from inadequate disease definitions balance out, and despite these biases, the disease association data still hold valuable insights into the

underlying pathomechanisms. In the fourth publication, we assessed the degree to which extent this implicit assumption is substantiated by data.

Assume d_1 and d_2 are two diseases sharing an unknown molecular mechanism M . If BEV network medicine uses data type D_1 containing information about d_1 and d_2 , for example disease-gene association data, M should result in significant similarities between d_1 and d_2 considering data D_1 . This translates to having an edge in the network G_1 that is constructed based on D_1 , where d_i 's are nodes. These networks are diseasesomes introduced earlier in this chapter. Similarly, for any other data types D_i used as input for the BEV approach, we expect to see similarity in the corresponding network G_i . Therefore, in the diseasesomes G_1 and G_2 generated based on similarities in D_1 and D_2 , the edge distribution should show a higher correlation than expected by chance in two random networks constructed under similar conditions.

Based on the BEV implicit assumption, we formulated two hypotheses to be tested:

1. The global-scale hypothesis suggests that networks formed from two distinct types of disease association data, both carrying valuable information about molecular mechanisms, should exhibit a higher level of pairwise similarity than what would be expected by chance.
2. The local-scale hypothesis proposes that this principle should apply not only on a global level but also within the specific neighborhoods of individual diseases represented as nodes in the diseasesomes.

To perform pairwise similarity analyses on diseasesome and drugome networks and their counterpart randomized networks, we used customized versions of graph edit distance (GED). GED is a measure used to quantify the dissimilarity between two graphs [199]. It is calculated as the minimum cost required to transform the source graph into the target graph using elementary edit operations, such as deleting, inserting, and substituting nodes and edges. To test the local-scale hypothesis, we defined a local version of GED (local node distances) which is calculated based on only the neighborhood of a node. GED is the only available graphs distance measure we are aware of that meets the criteria essential for our analyses. These requirements include: 1) a graph distance measure which is decomposable into local node distances, 2) the local node distances should depend on the node's local neighborhoods in the compared networks and not on the overall networks topologies, 3) due to node alignment between the compared networks, both the global network distance and the local distances should be node-identity-aware rather than permutation-invariant, i.e., node labels are important; and 4) for our large-scale permutation tests to be feasible, the distances must be computable in linear time relative to the size of the networks.

To test the global-scale hypothesis, we computed one-sided empirical P -values based on the comparison between randomized networks. For the local-scale

hypothesis, i.e., to test whether the local distances in the original networks are significantly smaller than the local distances in the randomized ones, we computed the one-sided Mann-Whitney U (MWU) test as well as node-specific local empirical P -values. Detailed explanation of the methods and rationale behind selecting GED as a network similarity measure for this study can be found in the original publication [131] (Appendix A.4).

4. Publications

4.1. Publication 1: Exploring the SARS-CoV-2 virus-host-drug interactome for drug repurposing

Citation

The following article titled “Exploring the SARS-CoV-2 virus-host-drug interactome for drug repurposing” has been published in Nature Communications on July 14, 2020.

Full citation:

Sepideh Sadegh[†], Julian Matschinske[†], David B. Blumenthal, Gihanna Galindez, Tim Kacprowski, Markus List, Reza Nasirigerdeh, Mhaned Oubounyt, Andreas Pichlmair, Tim Daniel Rose, Marisol Salgado-Albarrán, Julian Späth, Alexey Stukalov, Nina K. Wenke, Kevin Yuan, Josch K. Pauling & Jan Baumbach (2020). “Exploring the SARS-CoV-2 virus-host-drug interactome for drug repurposing.” *Nature communications*, 11(1), 3518; <https://doi.org/10.1038/s41467-020-17189-2>

[†] These authors contributed equally.

Summary

De novo drug development takes years to result in treatment for a disease. Through the identification of additional applications for already approved drugs, drug repurposing is a better alternative approach, particularly in the case of a new rapidly spreading pandemic, such as the recent COVID-19 outbreak. In the initial months after identifying SARS-CoV-2 virus in January 2020, there were numerous studies conducted with the aim of shedding light on the molecular mechanism underlying SARS-CoV-2 viral infection and consequently finding a treatment. One of the earliest and most important studies was done by Gordon et al. [65] where they could identify 332 high-confidence protein-protein interactions between the virus proteins and the human proteins. Additionally, there existed other valuable sources of molecular and drug data, along with various network medicine algorithms developed previously, that could be exploited for the task of COVID-19 drug repurposing. However, such information and algorithms were scattered across multiple publications and not available in an integrative fashion to facilitate drug repurposing studies.

We developed CoVex, an online platform available at <https://exbio.wzw.tum.de/covex/>, that aims to address this issue by providing an interactive resource for exploring the SARS-CoV-2 virus-host interactome and identifying potential drug targets. CoVex integrates information on virus-human

protein interactions, human protein-protein interactions, and drug-target interactions. It allows users to explore the virus-host interactome and use network-based algorithms to find potential drug targets and repurposable drug candidates. The inclusion of median gene expression levels, specific to each tissue from the GTEx data portal, facilitates the application of tissue-specific filtering to the results. The idea behind CoVex is to identify and target host proteins which are key modulators and necessary for the virus' life cycle, instead of targeting virus protein. These host key proteins are not necessarily the first interactors of virus proteins but downstream proteins in the human interactome indirectly affected by the virus proteins and can be identified by the network medicine algorithms integrated into the CoVex tool.

Presenting four different scenarios, we demonstrated the utility of CoVex. Users can guide the analysis by selecting (usually) hypothesis-driven starting points (viral proteins, human proteins or drugs), so-called seeds, and further leverage their expert knowledge in the subsequent steps. Based on the initial hypothesis and selected seeds, intended systems medicine algorithms find connecting paths from viral proteins to drugs using host proteins as proxies. The integrative and interactive online platform CoVex is intended to make COVID-19 drug research more accessible by helping researchers understand the molecular mechanisms of COVID-19, test their hypotheses, and prioritize candidate therapeutics.

Contribution

As stated in the publication: “S.S., J.M., J.B., M.L., T.K., J.K.P., A.P., and A.S. conceived and designed the study. S.S. and J.M. were in charge of overall direction, planning, and supervision. S.S., G.G., T.D.R., M.S.-A., and N.K.W. performed the acquisition, integration, and interpretation of data. S.S., D.B.B., M.L., and K.Y. developed and adapted the algorithms for network-based drug repurposing. J.M., R.N., M.O., and J.S. implemented the web platform. All authors provided critical feedback and helped in the interpretation of data, manuscript writing, and the improvement of the platform.”

In detail: I contributed to conceptualizing and designing the study. I had the leading role in systems and network medicine algorithms adaptation, integration and development of the new algorithm MuST, for network exploration, prediction of drug target and drug candidates. I had the main role in integration of data in the platform. I contributed to the writing, creation of figures, and revision of the manuscript and supplementary information.

Rights and permissions

The publication is available in Appendix A.1. "© The Author(s) 2020. Published by Springer Nature. This is an open access article under the Creative Commons CC BY

license (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited" [11].

Additional supplementary material

Supplementary data are available online at Nature Communications:
<https://doi.org/10.1038/s41467-020-17189-2>

4.2. Publication 2: Network medicine for disease module identification and drug repurposing with the NeDRex platform

Citation

The following article titled “Network medicine for disease module identification and drug repurposing with the NeDRex platform” has been published in Nature Communications on November 25, 2021.

Full citation:

Sepideh Sadegh[†], James Skelton[†], Elisa Anastasi, Judith Bernett, David B. Blumenthal, Gihanna Galindez, Marisol Salgado-Albarrán, Olga Lazareva, Keith Flanagan, Simon Cockell, Cristian Nogales, Ana I. Casas, Harald H. H. W. Schmidt, Jan Baumbach, Anil Wipat & Tim Kacprowski (2021). “Network medicine for disease module identification and drug repurposing with the NeDRex platform.” *Nature communications*, 12(1), 6848; <https://doi.org/10.1038/s41467-021-27138-2>.

[†] These authors contributed equally.

Summary

There is currently a significant challenge in traditional drug discovery due to a lack of effectiveness. Repurposing existing drugs for treatment of diseases other than their original indications can be a cost-effective and faster alternative. Among the variety of in silico methods, mechanistic drug repurposing approaches, i.e., targeting the mechanism underlying disease, have been advantageous due to the interpretability of their results leading to making informed decisions about potential candidates rather than dealing with predictions coming out of a black box that cannot be interpreted.

Previous research suggests that genes associated with diseases are not randomly distributed within biological networks; instead, they exhibit a tendency to cluster together in what are referred to as *disease modules* – small interconnected subnetworks representing mechanisms that can be related to the phenotype. One of the fundamental principles of network medicine is the notion that diseases can be perceived as disruptions to these modules. Accordingly, in silico drug repurposing can be done by, firstly, discovering such disease modules and subsequently finding drugs targeting genes or proteins within these modules. In order to do so, we need to build heterogeneous biological networks from pertinent data. However, this data is scattered across a multitude of databases, each using their own vocabulary for diseases.

Furthermore, the existing drug repurposing studies are either restricted to predicting treatments for particular diseases (of the interest of study) or developing non-translational algorithmic approaches. Computational drug repurposing methods often neglect valuable expert knowledge. By incorporating the concept of human-in-the-loop into drug repurposing, we can leverage the expertise of researchers in the field of pharmacology and biomedicine to assess the results at each stage of the workflow, resulting in more promising predictions.

NeDRex is a platform that aims to address the need for adaptable drug repurposing tools by providing an interactive and integrative tool, enabling biomedical researchers to employ network-based approaches for drug repurposing and disease module discovery of their individual use cases while benefiting from the notion of human-in-the-loop. NeDRex integrates data from different sources, including genes, drugs, drug targets, and disease annotations. It allows users to construct biological networks, mine them for disease modules, prioritize drugs targeting disease mechanisms, and perform statistical validation.

The NeDRex platform is built of three main components: NeDRexDB knowledge base (available at <http://neo4j.nedrex.net/> and <https://api.nedrex.net/>), NeDRexApp (a Cytoscape app as the user interface of the platform to run the algorithms implemented in the backend, available at <https://apps.cytoscape.org/apps/nedrex>), and NeDRexAPI (the RESTful API, to give access to the knowledge base available at <https://api.nedrex.net/>).

The utility of NeDRex is showcased through five distinct use cases, with results assessed using implemented statistical methods. These use cases include: 1) identification of disease pathways for ovarian cancer; 2) identification of therapeutic drugs for inflammatory bowel disease; 3) drug and drug target identification for pulmonary embolism; 4) identification of disease module and drug candidate for Huntington's disease; and 5) hypothesis-driven drug repurposing for Alzheimer's disease.

Contribution

As stated in the publication: “S.S., J.S., D.B.B., J.Ba., A.W., and T.K. conceived the idea and designed the platform. S.S., J.S., J.Be., E.A., G.G., K.F., S.C., T.K. performed the acquisition, harmonization and integration of databases. S.S. and D.B.B. developed and adapted the network-based algorithms for drug repurposing. S.S., E.A., G.G., M.S.-A., O.L., C.N., and A.I.C. discovered and approved the use cases. J.S. implemented the API. S.S. and J.Be. implemented the Cytoscape app. All authors provided critical feedback and discussion, assisted in the interpretation of data and use cases, writing the manuscript, and the improvement of the platform.”

In detail: I significantly contributed to conceptualizing the idea and designing the platform. J.S. and I conducted the data acquisition, harmonization and integration

into the platform. I implemented all the network medicine algorithms for disease module identification, drug prioritization, and statistical validations and further developed my previous version of MuST (from CoVex) for the NeDRex platform. I implemented the front-end of the platform as a Cytoscape App. I discovered and validated most of the use cases. I greatly contributed to the writing and revision of the manuscript and supplementary information. I generated all the figures.

Rights and permissions

The publication is available in Appendix A.2. "© The Author(s) 2021. Published by Springer Nature. This is an open access article under the Creative Commons CC BY license (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited" [12].

Additional supplementary material

Supplementary data are available online at Nature Communications:
<https://doi.org/10.1038/s41467-021-27138-2>

4.3. Publication 3: Lessons from the COVID-19 pandemic for advancing computational drug repurposing strategies

Citation

The following article titled “Lessons from the COVID-19 pandemic for advancing computational drug repurposing strategies” has been published in Nature Computational Science on January 14, 2021.

Full citation:

Gihanna Galindez†, Julian Matschinske†, Tim Daniel Rose†, **Sepideh Sadegh†**, Marisol Salgado-Albarrán†, Julian Späth†, Jan Baumbach & Josch Konstantin Pauling (2021). “Lessons from the COVID-19 pandemic for advancing computational drug repurposing strategies.” *Nature Computational Science*, 1(1), 33-41; <https://doi.org/10.1038/s43588-020-00007-6>.

† These authors contributed equally.

Summary

In contrast to traditional drug discovery, repurposing approved drugs offers a swift and efficient method to discover alternative treatments by identifying new applications for medications with established safety and pharmacological profiles. Particularly in the context of the COVID-19 pandemic that rapid and cost-effective approaches are essential, *in silico* drug repurposing is the method of choice to identify novel treatments. Therefore, there were many efforts during the COVID-19 pandemic on drug repurposing.

In this review, we summarized the methodology used in COVID-19 drug repurposing research, discussed the challenges encountered and lessons we learned from them. First, we gathered and reviewed the data sources used in COVID-19 drug repurposing studies, including molecular data resources, networks and interaction resources, drug databases, and clinical trial resources. In the next step, we grouped computational methods into two general virus-targeting and host-targeting approaches.

To find potential inhibitors for viral proteins, most virus-targeting strategies utilized structure-based drug screening techniques by using docking simulations. Unlike conventional docking protocols that are constrained to analyzing millions of chemical compounds, deep learning approaches, such as neural networks, have the capacity to analyze billions of compounds and have been increasingly used for COVID-19 drug screening.

Host-targeting strategies entail the identification of potential drugs that disrupt host mechanisms involved in viral pathogenesis, therefore, advantageous as they are less susceptible to drug resistance. These approaches involve integration and analysis of multiple omic data types and either use data-driven network-based methods or signature-based methods. The latter involve the discovery of drug-induced expression profiles that demonstrate contrasting patterns to the signature observed in COVID-19.

In this review work, we also compared the predictions of the computational methods to the drugs undergoing clinical trials available at the time of the publication. Finally, we highlighted the lessons learned from the reviewed drug repurposing efforts, and proposed a unified drug repurposing strategy to improve readiness in the event of future outbreaks. The unified strategy included standardizing molecular databases, combining host- and virus-targeting approaches, synergistic drug combination, combining computational and experimental research, inclusion of expert-guided analyses, and validation of candidates.

Contribution

As stated in the publication: “G.G., J.M., T.D.R., S.S., M.S.A., J.S., J.B. and J.K.P. contributed equally to the manuscript writing. J.B. and J.K.P. were in charge of overall direction, planning and supervision. All authors provided critical feedback and helped to improve the manuscript.”

In detail: I focused, with G.G., on collecting and reviewing COVID-19 drug repurposing studies using host-targeting approaches and comparison of their predictions to clinical trials. I contributed with J.M. and J.S. to the collection of “Data resources” used in COVID-19 studies. Together with all other co-authors, I equally contributed to the “Lessons learned” section, generation of figures, writing and revision of the manuscript and supplementary information.

Rights and permissions

The publication is available in Appendix A.3 with permission of Springer Nature. "© Springer Nature America, Inc. 2021. Ownership of copyright in this original research article remains with the Author, and provided that, when reproducing the contribution or extracts from it or from the Supplementary Information, the Author acknowledges first and reference publication in the Journal, the Author retains the right to reproduce the contribution in whole or in part in any printed volume (book or thesis) of which they are the author(s)" [13].

Additional supplementary material

Supplementary data are available online at Nature Computational Science:
<https://doi.org/10.1038/s43588-020-00007-6>

4.4. Publication 4: Lacking mechanistic disease definitions and corresponding association data hamper progress in network medicine and beyond

Citation

The following article titled “Lacking mechanistic disease definitions and corresponding association data hamper progress in network medicine and beyond” has been published in Nature Communications on March 25, 2023.

Full citation:

Sepideh Sadegh, James Skelton, Elisa Anastasi, Andreas Maier, Klaudia Adamowicz, Anna Möller, Nils M. Kriege, Jaanika Kronberg, Toomas Haller, Tim Kacprowski, Anil Wipat, Jan Baumbach, David B. Blumenthal (2023). “Lacking mechanistic disease definitions and corresponding association data hamper progress in network medicine and beyond.” *Nature Communications*, 14(1), 1662; <https://doi.org/10.1038/s41467-023-37349-4>.

Summary

One of the network medicine’s goals is to find better treatment for diseases that are not only palliating symptoms. To achieve this, a shift in disease definition from organ- and phenotype-based to mechanism-based is essential. Targeting distinct molecular mechanisms underlying the disease, so-called endotypes, can actualize disease-modifying treatments.

Close-up network medicine studies focus on a specific disease and conduct their analyses with molecular data obtained from well-characterized patient cohorts, while bird’s-eye-view (BEV) network medicine approaches use large-scale disease association data often gathered from a multitude of data sources. The BEV approaches look at diseases and the relationships between them in a more global view and are not intended to focus only on a specific disease.

Some sources of bias in the data used by BEV approaches have been studied before. They include but are not limited to the effect of incompleteness of disease-gene association and protein-protein interaction (PPI) data, as well as potential bias originating from highly studied proteins and genes. While biases from the non-disease side (like genes) in disease association data are well-explored, potential biases from the disease side itself are often overlooked. One of these biases is how in large-scale disease association databases, diseases are currently annotated with the very phenotype-based disease definitions that the network medicine field aims to transition from. Therefore, BEV approaches that rely on such disease association data run the risk of systematically replicating the biases introduced by these disease definitions. As a result, BEV approaches operate under the implicit

assumption that the biases arising from phenotype-based disease definitions balance out, and despite these biases, the disease association data obtained using such definitions still hold valuable insights into the underlying pathomechanisms yet to be discovered.

To assess to what degree the aforementioned underlying assumption is supported by data, we first constructed disease–disease networks based on different types of disease association data (such as gene, comorbidity, symptom, and drug) as well as drug–drug network based on indication and drug target data. Then, we performed two types of similarity analysis in global- and local-scale to compare the constructed networks and tested the hypotheses arising from the implicit assumption of BEV network medicine. We implemented the similarity analyses in the Python package graph similarity quantification tool (GraphSimQT), enabling users to quantify biases originating from other disease association data types other than the ones covered in our study. The GraphSimQT package is available at: <https://github.com/repotrial/graphsimqt>. The results are explorable via the web interface graph similarity visualizer (GraphSimViz) <https://graphsimviz.net>.

The results indicate strong evidence in favor of the global-scale hypothesis. However, they only offer limited support for the local-scale hypothesis. The interpretation of results is that BEV network medicine offers only a distant perspective on the yet-to-be-uncovered endotypes. When we closely examine individual diseases, the clarity of the picture diminishes that can be due to the fuzzy disease definitions. These findings imply that network medicine approaches should not blindly rely only on general publicly available disease association data, which uses ill-defined disease classifications, and need to be complemented with additional layers of molecular data for well-characterized patient cohorts to be used for close-up analyses.

Contribution

As stated in the publication: “D.B.B. and S.S. conceived and designed this study and implemented the GraphSimQT Python package to compare the different networks. S.S. carried out the analyses. J.S., S.S. and K.A. integrated the data and constructed the networks. A.Ma. implemented the GraphSimViz web tool. D.B.B., S.S., E.A., N.M.K., and A.Mö. drafted the manuscript. J.K., T.H., and the Estonian Biobank Research Team provided the comorbidity data. D.B.B., J.B., and A.W. supervised the project. All authors provided critical feedback and discussion and assisted in interpreting the results and writing the manuscript.”

In detail: I, together with D.B.B conceptualized the idea and designed the study. I had the main role in data acquisition, integration, and network generation. J.S. and I conducted the mapping between disease vocabularies and other necessary data

harmonization steps. I constructed the comorbidity network based on the data acquired from the Estonian Biobank. I, together with D.B.B implemented the network comparison algorithms as a Python package (GraphSimQT: graph similarity quantification tool). I performed all the analyses and plotted the results. I contributed to the writing and revision of the manuscript and supplementary information. I generated all the figures.

Rights and permissions

The publication is available in Appendix A.4. "© The Author(s) 2023. Published by Springer Nature. This is an open access article under the Creative Commons CC BY license (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited" [131].

Additional supplementary material

Supplementary data are available online at Nature Communications:
<https://doi.org/10.1038/s41467-023-37349-4>

5. General Discussion and Outlook

Systems medicine is an interdisciplinary and integrative approach that extends the concept of systems biology by applying computational methods with the aim of deciphering complex biological systems. This, in turn, leads to novel clinically relevant applications, contributing to the development of more effective prognostic, diagnostic, and therapeutic strategies [200]. As our knowledge of biological processes expanded, we learned that various components- such as proteins, RNA, DNA, metabolites, and environmental factors- engage in intricate interplay with one another. Therefore, modeling cellular interactions as networks has emerged as a natural approach, leading to the growing importance of network biology in bioinformatics research. Furthermore, other types of networks from non-molecular origin, e.g., from population-based patient clinical histories, which investigate diseases from another perspective, can build on the knowledge gained from molecular-based networks. By constructing diverse networks presenting different relationship types and applying principles of graph theory to analyze them, network medicine, as an offshoot of systems medicine, uncovers complex disease mechanisms and identifies potential therapeutic targets that may be missed by traditional approaches.

In the course of my PhD, I developed two integrative and interactive drug repurposing platforms, one specifically for COVID-19 and the other for general medical conditions, employing network medicine methods. The aim was to leverage network medicine algorithms in combination with a comprehensive and harmonized knowledge base enriched with relevant disease and drug data, to develop interactive platforms that are adaptable to biomedical researchers' individual use cases. Such a platform enables the researchers to study mechanisms underlying diseases and consequently to find potential repurposable drugs.

In this dissertation, I also presented a review on the methodologies used in COVID-19 drug repurposing research, discussing the existing challenges, highlighting the lessons learned from them, and proposing a unified drug repurposing strategy to improve readiness for future outbreaks.

During the writing of one of the review works for my PhD [16], I learned that in the broad domain of systems medicine, various studies use diverse types of data, many integrated from large-scale disease related data. However, how the diseases are defined varies a lot across different data sources. This together with the challenges we faced to build a harmonized knowledge base brought me to the idea of exploring disease definitions further. Therefore, as the second main goal of this dissertation I assessed a less explored type of bias that disease-associated data introduces to network medicine approaches, that is how we currently define diseases.

In the subsequent discussion, while briefly going through all presented publications in this dissertation, I will discuss the limitations of our work and the

challenges that in general the field encounters. Lastly, I will explore the outlook for the drug repurposing domain.

5.1. COVID-19 drug repurposing with CoVex

With CoVex, we aimed to develop a web tool during the first months of the Covid-19 pandemic when there were no harmonized data resources that coalesced different types of useful information for the purpose of finding a treatment. We integrated SARS-CoV-2 virus–human interactions derived from the early AP-MS study of Gorden et al. together with publicly available human interactome (PPI networks) and drug–target information in a unified network with the focus on unfolding novel drug targets downstream in the human interactome instead of finding drugs targeting directly viral proteins. I adapted a selection of established network-based analysis methods and implemented a new one based on the Steiner tree concept to be employed in the network medicine context for drug target identification and drug prediction. These algorithms enable biomedical researchers to explore SARS-CoV-2 virus–host–drug interactome and test their hypotheses while allowing them to capitalize on their knowledge to guide the analysis. We presented a range of application scenarios based on the type of starting points of analysis like viral protein, human protein, drugs, or combination of them and showed how using different network-based algorithms implemented in the tool can provide insights into the molecular mechanisms of SARS-CoV-2 pathogenicity and predict potential repurposable drugs which were at the time of study undergoing clinical trials. For instance, we could identify proteins, which play a role in virus host cell entry and can be targeted by ACE inhibitors that were widely used in clinical trials to treat COVID-19. We could extract a potential immune-related mechanism triggered by the virus, and could consequently predict drugs targeting the mechanism which were being assessed in clinical trials. The network medicine approach used in CoVex for SARS-CoV-1 and SARS-CoV-2 is easily extendable to other viruses by integrating the corresponding virus–human interaction data.

Similar to other *in silico* methods, CoVex can merely suggest putative drug candidates for further investigation. Although the proposed drugs aim at proteins involved in a supposedly important mechanism for the virus, their actual impact needs confirmation through subsequent investigations. Inhibiting a cofactor that normally prevents the virus from manipulation of host proteins could potentially serve in favor of the virus. One of the limitations of CoVex is that selecting the algorithm and its associated parameters is not a simple task which can lead to lengthy explorative analyses. We tried to mitigate this to some extent by enabling parallel execution of multiple analyses. Another limitation is that different sources of drug–target interactions were treated similarly in the constructed network, while the strength of experimental evidence may vary depending on the utilized

experimental assay. Given that the integration of virus-human interaction data relied on hastily conducted research, enhancing the quality of the data could be achieved through subsequent updates, revisits, and manual curation. While data on gene expression levels per tissue type was included for post-filtering the results, it could be more advantageous to develop and integrate algorithms into the platform that utilize this information for predicting drug targets.

5.2. Interactive and integrative drug repurposing with NeDRex

To our knowledge, research in the domain of *in silico* drug repurposing has been limited to either non-translational algorithms or disease-specific predictions. Consequently, there exists an ongoing demand for comprehensive tools that enable non-computer scientists and biomedical researchers to easily employ computational drug repurposing methods and tailor them to their unique use cases. Among the variety of *in silico* drug repurposing methods, those designed to target the underlying mechanisms of diseases appear promising. This is because clinicians can interpret the results, as opposed to merely evaluating predictions derived from methods that function as black boxes.

As mentioned earlier in this dissertation, networks provide an effective method for representing biological data and elucidating the connections among diverse molecular entities. Our NeDRex platform employs a systems medicine approach which uses networks as a presentation model for data and graph-based methods to mine these networks in order to firstly uncover disease-related mechanisms, i.e., disease module identification (DMI) step, and subsequently find drugs targeting those mechanisms, i.e., drug prioritization step. The first challenge in developing this platform was to build a harmonized knowledge base from the plethora of databases which were relevant for the purpose of drug repurposing. Building such a unified knowledge base requires mapping different systems of identifiers into one system. This is an easier task for entities like drugs and genes where identifiers are well defined. The mapping task becomes particularly difficult when dealing with databases containing disease-related information since each uses a different disease vocabulary system. After investigating available disease vocabulary options, we opted for MONDO as the target vocabulary which has the highest mappability for disease vocabularies used in the databases we had aimed to integrate into the NeDRex knowledge base. Moreover, it benefits from a hierarchical structure capturing both disease umbrella terms and more specific sub-types. The second challenge was the selection of the best performing network-based method for the task of DMI. Since there is not a single method outperforming the others and the most suitable strategy for DMI also depends on the specific application at hand [152], we integrated several algorithms in the

platform with different underlying paradigms to provide specific exploration options for various particular research questions and hypotheses. All these algorithms use prior knowledge on PPIs in the form of a network and range from an unsupervised method using ant colony optimization for data-driven patient subgrouping [44] to a proximity-based method using Steiner trees for detecting molecular pathways underlying diseases [11]. We developed three statistical validation methods to evaluate the outcome of DMI and drug ranking analyses, individually and combined, which give users a measure for comparison of the outcome of different algorithms applied to their specific use case. Five use cases have been showcased, illustrating NeDRex's capacity to extract biologically relevant candidate disease modules and potentially repurposable drugs. Specifically, we demonstrated how NeDRex could identify promising drugs worthy of further exploration for the treatment of inflammatory bowel disease, pulmonary embolism, Huntington's disease, and Alzheimer's disease.

Although the expert-in-the-loop paradigm is a significant advantage of the NeDRex platform, it also represents its primary limitation. When utilizing NeDRex, investing domain knowledge is not merely an option but a necessity. Without leveraging this expertise, the likelihood of obtaining biologically meaningful disease modules or promising drug repurposing candidates is diminished. Crucially, even the devised statistical evaluation methods based on empirical *P*-values cannot substitute for the expert user, as they, too, rely on existing knowledge and drugs undergoing clinical trials. Ultimately, the NeDRex platform is not exempt from the limitations present in the integrated databases. These constraints encompass incompleteness of available PPI data, false positive PPIs, literature bias arising from over- and under-studied genes, and the lack of discrimination between activation and inhibition in the drug-protein associations available in the integrated databases. A corrective measure to mitigate literature bias will be introduced later in the Limitations and challenges section.

5.3. Lessons from COVID-19 to improve computational drug repurposing strategies

Our review on the methods used in COVID-19 drug repurposing research showed that most of the studies did not perform experimental evaluation. Therefore, to assess the quality of computational predictions, we checked the intersection between the final predicted drug lists of the individual studies and the drugs undergoing clinical trials. Additionally, for the virus-targeting approaches, we collected *in vitro* screening data, including IC₅₀ values for viral targets and inhibition indices from cell culture studies for SARS-CoV-2. Even though some predicted drugs were already used for the treatment of clinically ill COVID-19 patients, in the course of this review work, we noticed that drugs that advanced to

later stages in clinical trials were not chosen through *in silico* predictions but rather repurposed based on the clinical experience gained from previous outbreaks such as SARS or MERS. Their selection was primarily influenced by their known effects in palliating disease symptoms. Additionally, the majority of the reviewed studies did not proceed to experimental validation of the predictions. This translational disconnect between computational efforts for drug repurposing and their practical application in clinical settings represents a significant and universally recognized impediment in both drug repurposing and the field of medicine. The outcomes of systematic validation efforts will play a crucial role in identifying algorithms and datasets particularly suitable for drug repurposing in the context of COVID-19. The close collaboration among clinicians, experimental biologists, and computational biologists is needed to bridge this gap effectively. Since we are in need of computational tools that are able to deliver promising repurposable candidates, which in return could be validated in clinical trials, we proposed a unified strategy which is necessary for a more effective computational drug repurposing pipeline. Most parts of this unified strategy is not limited to the viral pandemic cases and can be slightly adapted to be applicable for a general disease condition. Our suggested strategy that also improves our readiness for future outbreaks has the following six main components:

1. **Standardized databases:** Building such a database is the key initial step of the pipeline. Some part of the required data containing general information, for example about drugs, proteins, and other diseases, is common in drug repurposing for any viral disease. Newly developed computational methods often use the same existing types of data. Therefore, establishing standardized databases which can be just upgraded with new virus data is of paramount importance. In the context of improving research quality, it is noteworthy that experimental replication of datasets obtained from different laboratories is an important step to enhance robustness. Existing drug clinical trial sources were employed in the development of some drug repurposing pipelines. Nevertheless, the clinical trial resources lacked standardization, posing challenges in analyzing trials for specific drugs due to variations in names, spellings, or typographical errors which could be avoided by using standardized drug IDs.
2. **Tool accessibility:** Method accessibility in the form of user-friendly tools allows researchers to run custom analyses using the developed algorithms, for example, on newly obtained experimental data. This makes it more likely that non-computer scientists and clinicians use these tools and continue with validation routines, leading to accelerating research. Despite the diverse methodological studies for COVID-19 drug repurposing, the availability of the interactive tools that clinicians can use was very limited.
3. **Consolidation of predictions (ensemble methods):** There were a few studies that combined outputs of different drug repurposing models or aggregated

results from different algorithms. These studies exhibited the highest proportion of overlaps with drugs that underwent clinical trials. This indicates that by consolidating various approaches, there is potential to substantially enhance confidence in repurposed candidates and offer valuable guidance to clinical researchers during the drug selection process. To this end a streamlined solution is necessary that encompasses tool accessibility and standardization, such as a centralized database storing drug candidate predictions, facilitating meta-analyses.

4. Development of combinatorial treatment: The computational identification of synergistic drug combinations is a relatively unexplored area that holds immense potential to enhance clinical decision-making. This approach has shown to be more effective than discovering individual monotherapies [201,202]. Methods that try to identify complementary drug groups and simultaneously take into account side effects are currently lacking. In the context of viral disease, combining drugs targeting the virus with the ones targeting the host is a promising strategy because of the simultaneous blocking of viral entry into cells and disrupting disease progression by inhibiting viral replication and host pathways. In vitro evaluation of thousands of compounds is manageable, but validating combinations is more challenging. Predicted combinatorial treatments could significantly narrow down the search space for later in vitro validation. The potential of existing screening databases, like the NIH OpenData portal [203] and the ReFRAME library [204], remains underutilized, but by integrating them with computational predictions, they could bridge the gap between in silico and in vitro research, facilitating the identification of promising combinatorial treatments.
5. Expert-guided analysis: Due to the limited understanding of the complex biological mechanisms behind COVID-19, expert knowledge or manual curation has been necessary at various pipeline steps, such as protein or pathway selection and drug prediction filtering. Expert screening aims to uncover inconsistencies or contradictions while enabling the discovery of new predictions, playing a vital role in filtering candidate drugs to identify potential adverse side effects. Hence, close collaboration between computational and clinical researchers becomes crucial, given the current limitations of computational approaches in terms of information on side effects and drug inhibitory or activatory actions on the targets.
6. Candidate validation strategies: Typically, drug repurposing studies validate their computational models by establishing their own "ground truth", which may comprise in vitro screening of predicted drugs, in vivo tests with animal models, ongoing clinical trials, literature mining, or expert knowledge [205]. As a result, there is significant diversity in the origins of these standards. There have been some endeavors to resolve this issue, such as databases like the NIH's OpenData portal, which aggregates and continually updates in vitro

screening data for thousands of compounds and other SARS-CoV-2-related assays. Such resources can be utilized for further validation or filtering of *in silico* predictions. In the course of our review, we found with the exception of one study [206], there has been no direct follow-up experimental validation in the drug repurposing endeavors for COVID-19. In the other reviewed studies, a variety of above-mentioned ground truths were used for the validation of predictions. Since it is impracticable to systematically validate all candidate drugs, leveraging the expert knowledge for assessing the predictions becomes more crucial. From the recent list of FDA approved drugs for COVID-19 treatment as of May 25, 2023, Ritonavir and Remdesivir were among the predictions returned by virus-targeting approaches. This shows the potential of computational drug repurposing methods to predict effective therapeutics, hence the importance of follow-up validations.

5.4. The bias introduced to network medicine by inadequate disease definitions

In many network medicine studies, data from public databases are used to infer relationships between diseases based on different types of commonality, such as shared symptoms, shared involved genes, and comorbidity. Public databases can also be utilized to link drugs to each other based on molecular similarity, functional similarity, and common targets. A multitude of network medicine studies have the bird's eye perspective towards the tasks of drug repurposing and uncovering mechanisms driving diseases (we call this type of approach BEV methods). Subsequently they use relationships between diseases, drugs, as well as diseases and drugs derived from large-scale databases. In contrast, some network medicine studies have a close-up and per-disease perspective and use mainly patient cohorts' molecular data rather than focusing on the relationships between diseases. The studies adopting the latter type of approach have led to higher translational findings [207,208].

We postulated that the translational underperformance of BEV methods might be due to the bias introduced by using large-scale disease association data, where diseases are annotated with the current inadequate disease definitions that are mainly organ- and phenotype-based. Given that molecular mechanisms driving diseases are often still not known, the majority of disease names do not denote such mechanisms but are after the doctor's name who coined the disease term, affected organs or locations in the body, or symptoms of the disease. ICD-10 codes that is the standard vocabulary to use in clinical settings have overly inclusive designations, ranging from symptoms, to syndromes, to definable molecular causes.

This disease naming principle results in data that is fuzzy, as it does not distinguish between “true” diseases with separate pathomechanisms, but it combines them to one disease due to shared symptoms or affected organs. This blurriness also brings about significant clinical implications, such as patients with mechanistically different true diseases receiving the same broad treatment and not a targeted treatment because diagnosed with one fuzzy disease.

The network medicine field aims to transition from these fuzzy disease definitions by discovering molecular mechanisms underlying the disease phenotypes, so-called endotype. This implies that BEV approaches, that still use the data based on fuzzy definitions, assume that the biases arising from inadequate disease definitions balance each other out, and despite these biases, the disease association data still hold valuable insights into the underlying pathomechanisms. The secondary goal of my dissertation was to quantify to which extent this implicit assumption is actually supported by data. For this purpose, we formulated two testable hypotheses, in global- and local-scale based on GED (explained in detail in the General Methods chapter), that I investigated in the fourth publication. To test the hypotheses, I quantified the pairwise similarity of several diseasomes and drugomes constructed from different types of disease and drug association data.

The results of pairwise similarity analysis of diseasomes indicates that the global-scale hypothesis holds. In other words, the results provide substantial proof supporting the general legitimacy of the BEV network medicine framework. Nonetheless, the findings also suggest that the reliability of outcomes produced through the BEV network medicine methods diminishes when focusing closely on specific diseases, i.e., local-scale analyses indicate that the majority of comparisons of local distances for the original networks are not significantly smaller than the local distances for the permuted counterparts. The exception is the case where a very coarse disease vocabulary, closer to disease clusters than actual specific disease, is used in the networks’ construction. Consequently, our results affirm the issue of solely depending on data labeled with phenotype-based descriptions when aiming to reveal molecular pathomechanisms.

Databases containing disease-related data use different disease vocabularies. For example, OMIM uses OMIM terms, DisGeNET uses UMLS CUIs, and national health record databases containing diagnoses for patients normally use some version of ICD codes. These vocabularies vary in their levels of granularity and are created using diverse methods and for distinct intentions. For network medicine to collectively utilize disease associations from multiple data sources, it's necessary to map the data to a shared target vocabulary. During this mapping process, because of unmappable terms, some data is inevitably lost. Our analyses performed in three different target vocabularies with different levels of granularity, including MONDO IDs, UMLS CUIs, and ICD-10 3-character codes, show that the selection of disease vocabulary impacts the final results

considerably. With the higher analysis resolution, the significance of the obtained *P*-values decreases. MONDO and UMLS CUI have similar granularity levels and show similar results. When using ICD-10 3-character codes, which is coarser than former disease vocabularies and can be regarded as disease clusters or umbrella terms rather than individual disease subtypes, around half of all computed MWU *P*-values for local GEDs are significant at 0.001 level. When comparing the diseasomes as a whole using global GEDs, all empirical *P*-values are significant. Moreover, drugome comparisons have led to more significant results on a local level than diseasome comparisons. This finding confirms that in case of using well-defined underlying annotations, like drug vocabularies, the network-based analyses return more reliable results.

Our effort to detect discernible patterns among diseases, based on their small or large empirical *P*-values calculated using local GEDs, was not successful. This might stem from certain existing disease definitions aligning with true endotypes. We conjecture that in cases where our current definitions directly match endotypes, the hypothesis at the local scale remains valid. However, this cannot be evaluated since the neighborhood of a disease in a diseasome contributes to the GED values. Therefore, it is not possible to deconvolute the results based on single nodes.

Both gene-based and variant-based diseasomes were included in our comparison analyses. Disease-gene association data is mainly derived from disease variant information and we expect to see similar results for the networks constructed on the two association data types. However, local-similarity analyses indicate higher similarities between variant-based diseasome and other diseasomes than between gene-based diseasome and others. Despite the fact that different mutations in one gene can cause different disease phenotypes, this level of information cannot be conserved at the disease-gene level and is lost in the process of mapping variants to genes which is necessary to generate disease-gene association.

In sum, firstly, our results suggest that utilizing large-scale databases, which contain disease-related information, without careful consideration, as they depend on phenotype-based disease definitions, can be risky. Rather, we underscore that in network medicine, the preferred approach should be to focus on close-up approaches, wherein data scientists collaborate with biomedical researchers to collectively analyze both molecular and comprehensive phenotype data for the same individuals. Within this cooperative framework, a constructive feedback cycle can arise. Initial conjectures about disease subtypes and their underlying molecular mechanisms are shaped by analyzing molecular data. This understanding is then honed by incorporating extensive phenotyping (such as histological images and blood-derived biomarkers) and the expertise of clinicians. In the last step, the hypotheses find validation through preclinical investigations

like gain- or loss-of-function studies. Many studies following these strategies have yielded noteworthy insights into distinct disease mechanisms [4,209–212].

Secondly, a good way forward is unsupervised network medicine techniques that can not only identify potential pathomechanisms but also simultaneously categorize patients into distinct mechanistic subgroups without depending on potentially unreliable predefined phenotype-based subtype labels. Despite a limited number of such methods being available [44,213,214], the majority of current pathomechanism discovery approaches still hinge on contrasts between phenotypic cases and controls or gene lists linked to a (perhaps ambiguously defined) disease term [15,196,215].

Lastly, it's essential to highlight that the absence of well-defined mechanistic disease classifications not only impedes advancements in network medicine but also adversely impacts nearly all data-driven methods such as treatment formulation and diagnosis. These methods heavily depend on disease association data linked to phenotype-based disease descriptions. For example, an AI model designed to assist in diagnosis, trained using genetic disease signatures, will generate inaccurate outcomes if the disease labels employed during training do not accurately match the authentic endotypes.

Our method has the limitation that our findings cannot definitively eliminate the possibility that factors beyond inadequate disease definitions could contribute to the observed blurriness in the BEV network medicine at a local level. One of the potential factors introducing bias might be off-target effects of drugs that could impact the drugomes' analysis. Another potential source of bias is any gene-related data (explained in detail in the Limitations and challenges section) which can affect the analyses of the gene-based diseaseome. It is important to emphasize that the distributions of the derived local empirical P -values show the outcomes we obtained from all the analyses exhibit remarkable consistency across all utilized data modalities. Given that disease definitions are the sole confounders impacting all data categories, this provides substantial (albeit not definitively conclusive) indication that the observed blurriness in the local context can predominantly be linked to these definitions.

5.5. Limitations and challenges

Some of the limitations of the presented works in this dissertation are already discussed in the corresponding discussion section of each publication. Here, I mainly discuss the general limitations and challenges computational network-based drug repurposing approaches face.

Important biases in the data that network medicine approaches use have been reported and their effects have been studied. These biases include incompleteness of disease-gene association and PPI data [151], as well as study bias. Cancer-

associated proteins have higher degree in PPI networks due to being more investigated in multitude of studies, this can be one of the sources of study bias [216–219]. Patterns of publications on human genes are highly skewed [220]. It is estimated that a relatively small proportion of human genes receive a significant amount of research attention. This variation in publication numbers per gene may arise from the influence of past research priorities, which continue to shape present endeavors [221]. In other words, it is the consequence of the positive feedback loop or “the rich become richer”. Study bias goes beyond this and affects functional gene annotation resources [222] such as the Gene Ontology (GO) as well. These biases have several adverse consequences. It has been found that many DMI methods using PPI networks suffer from hub node bias and their models predominantly learn from node degrees rather than capitalizing on the biological information embedded within the network edges [223].

To mitigate literature bias, although we have introduced the hub penalty to our MuST algorithm as a degree correction measure to account for hub nodes in PPI networks (proteins with high degree in networks), the selection of the extent of penalizing depends on the user and not based on any prior information. This can be further improved by including the knowledge from databases such as IntAct where the numbers of times proteins have been used as bait and prey in AP-MS and Y2H studies are provided. This information can then be used to incorporate bias-aware edge costs in the PPI networks by making edges towards highly studied proteins more expensive. For other DMI methods, similar degree-correction solutions can be integrated into the method.

The current, constrained information on PPIs obtained through experimental approaches can be enhanced by modern deep learning algorithms' capable of predicting protein structures and PPIs. For example, protein 3D structure predicted by AlphaFold [224], based solely on its amino acid sequence, could be leveraged for predicting PPIs, reducing study biases, and supplementing the existing PPI networks. However, prediction methods have their own caveats. Predicted structures may not account for post-translational modifications (e.g., phosphorylation, glycosylation) that can significantly impact PPIs. Predicting interactions in protein complexes or multimeric proteins is more challenging than predicting interactions between individual proteins. Some PPIs are transient, while others are stable. Some involve weak, non-specific binding, while others are highly specific. Predicting the full spectrum of interactions accurately is a significant challenge. To predict PPIs, it's not enough to consider proteins as static structures. Their dynamic behavior, including conformational changes, flexibility, and the kinetics of interactions, must be taken into account. Computational methods that aim to predict PPIs need to incorporate dynamic modeling techniques to capture these aspects accurately. Failure to do so can lead to inaccurate predictions of how and when proteins interact in biological systems.

Using PPI networks, even the ones built based on experimental validation, as prior knowledge for the DMI task, comes with some inherent simplifications. These simplifications arise from several factors [2]:

- Temporal abstraction: PPI networks typically represent binary interactions, omitting precise timing and kinetics of signaling events that occur over various timescales.
- Spatial oversimplification: While signaling is spatially dependent, i.e., post-translational modifications and interactions depend on physical proximity between proteins, PPI networks lack explicit spatial information, failing to capture physical proximity's impact on interactions and subcellular compartmentalization.

One limitation of methods using gene expression as a proxy for protein activity is that it has been shown in some studies that gene expression can be misleading for measuring protein activity under some conditions [32–34].

Ideally, the evaluation of DMI and drug repurposing methods like any other computational approach would need gold standard datasets. However, for biomedical problems defining or generating gold standards is not feasible since our knowledge is not complete about any diseases. This makes the evaluation of drug repurposing very challenging. One workaround is to use proxies to establish indirect metrics for evaluation of methods. However, proxies cannot reflect the complexity of the real-world biomedical problems in its whole. Different validation strategies for computational drug repurposing have been used. The predominantly used ones in studies are: 1) case studies, 2) overlap of predictions with known drug indications and 3) sensitivity- and specificity-based methods [225]. The latter validation strategy, also the most rigorous type, requires using true negatives (failed drugs) and false positives. True negative set contains the drugs which have been tested for the treatment of a disease and did not pass the approval requirements. Although there have been some efforts like *repoDB* [226] to compile a list of unsuccessful indications from clinical trial databases, this list is based on only tested drug-indication pairs and far from being thorough enough to be used in a reliable validation approach. Current studies using sensitivity- and specificity-based validation methods consider all unannotated drug-indication pairs, i.e., the pairs that do not exist in publicly available databases, to be false positives. Labeling unannotated pairs as false implies that all emerging repurposing ideas are considered false positives. This contradicts the purpose, as computational repurposing methods aim to propose new indications, where no annotated connection exists based on the current data [225].

For our NeDRex platform, we implemented a validation method using empirically computed *P*-values based on the overlap of the predictions with a reference list, which consists of known drug indications and the ones undergoing clinical trials. We stress that if the list of reference drugs is not exhaustive or includes numerous

false positives, the P -values could lead to deceptive conclusions. As a result, the P -values are reliant on existing knowledge and, therefore, cannot replace but rather aid the expert in the loop.

Another challenge that network medicine faces is the incomplete mapping between different disease vocabularies. To make the best use of disease associations across diverse data sources, while each utilizing different disease terminologies as identifiers, it's necessary to map the data to a shared target vocabulary. The comorbidity data (population-based data) is reliant on ICD-10 codes, which have a very non-homogenous designation, and often not enough fine-grained. Therefore, mapping from ICD codes to other finer-grained disease vocabularies will introduce a lot of noise to the data. On the other hand, mapping other disease-associated data with finer-grained terminologies to ICD-10 codes reduces the specificity. This is a known and big challenge if we want to integrate epidemiological data with other disease-related data.

One of the challenges encountered during the data integration is that many databases, despite being publicly accessible, are either wholly or partially not open data. In other words, we are unable to publicly distribute them through a unified knowledge base, and the platform utilizing this data cannot have an open license. This limitation hinders both the reproducibility of results and the ability to contribute effectively to the advancement of science.

5.6. Conclusion and outlook

The NeDRex platform is under constant development and improvements both in terms of data and algorithms. Therefore, its current status differs from the version presented in this dissertation. GO terms as annotations are added to protein data as well as types of tissue that proteins are expressed in. Moreover, drug side effects information is integrated into the knowledge base allowing drug repurposing investigation based on disease symptoms versus drug side effects, which are basically the same. An improved and more robust version of MuST is also implemented in the updated platform. Inclusion of databases like repoDB which contain previous failed repurposing attempts will be a good addition assisting to filter the list of predicted drugs. Databases providing drug clinical trials could be as well a valuable addition for the evaluation of predictions. However, such databases normally do not use any conventional disease vocabularies in their system, hence, are very difficult to integrate in the knowledge base systematically.

As discussed earlier, it is tricky to choose which network-based DMI method to apply to the problem at hand. Some studies have shown meta or ensemble analysis using multiple methods increase the accuracy of drug predictions [206]. Such a method, for instance, either combines results from different algorithms into

concatenated modules, returns the consensus of different methods' outputs, or aggregates ranks of different methods' predictions.









We opted for implementing the front-end user interface of our drug repurposing platform as a Cytoscape app since it is a powerful tool to visualize biological networks and very popular among biomedical researchers with limited computational expertise. The NeDRexApp has been downloaded more than two thousand times as of May 2023. However, web tools running in browsers have advantages over stand-alone apps, namely, cross-platform compatibility, self-updating, less memory and computation load on system, no need to download and local installation. Therefore, we decided to extend the NeDRexApp to a web tool and later even went further to develop a web-based plugin enabling standardizing and simplifying network analysis as well as facilitating visual exploration of networks for any biomedical web tools. This idea led to Drugst.One [227] (under development), a plug-and-play solution for online network medicine drug repurposing which is also coupled with an updated and extended version of the NeDRexDB knowledge base. By using this customizable plugin, any web tools' results can be visualized in the standardized format. It also enables users to run further analyses using integrated network medicine algorithms or explore their results while adding information from the knowledge base.

The four publications included in this thesis are part of the large EU-funded project REPO-TRIAL and are a subset of all activities that were carried out in that context. Since the onset of REPO-TRIAL, the computational biology team of the consortium sought to 1) integrate relevant databases to construct heterogeneous networks to mine for disease module identification and drug repurposing and 2) develop computational tools that contribute to a deeper understanding of the molecular mechanisms underlying diseases. As one of the outcomes of these efforts, drugs repurposed *in silico* for specific diseases (selected within the framework of the REPO-TRIAL project) are undergoing validation in clinical studies. The continuation of REPO-TRIAL emerged as another EU-funded project, called REPO4EU, the overarching objective of which is to establish and expand an online platform, providing validated precision drug repurposing on a global scale. In future, more endeavors like REPO-TRIAL and REPO4EU are needed to make drug repurposing as the new gold standard in drug development. To this end, not only research but also regulatory measures come into play. This work is only a stepping stone and with improved data availability, collaboration, applied clinical trials, platforms like NeDRex and similar will further improve in quality and impact of results.

A. Appendix

A.1. Exploring the SARS-CoV-2 virus-host-drug interactome for drug repurposing

Exploring the SARS-CoV-2 virus-host-drug interactome for drug repurposing

Sepideh Sadegh^{1,6}, Julian Matschinske^{1,6}, David B. Blumenthal¹, Gihanna Galindez¹, Tim Kacprowski¹, Markus List¹, Reza Nasirigerdeh¹, Mhaned Oubounyt¹, Andreas Pichlmair², Tim Daniel Rose³, Marisol Salgado-Albarrán^{1,4}, Julian Späth¹, Alexey Stukalov², Nina K. Wenke¹, Kevin Yuan¹, Josch K. Pauling³ & Jan Baumbach^{1,5}✉

Coronavirus Disease-2019 (COVID-19) is an infectious disease caused by the SARS-CoV-2 virus. Various studies exist about the molecular mechanisms of viral infection. However, such information is spread across many publications and it is very time-consuming to integrate, and exploit. We develop CoVex, an interactive online platform for SARS-CoV-2 host interactome exploration and drug (target) identification. CoVex integrates virus-human protein interactions, human protein-protein interactions, and drug-target interactions. It allows visual exploration of the virus-host interactome and implements systems medicine algorithms for network-based prediction of drug candidates. Thus, CoVex is a resource to understand molecular mechanisms of pathogenicity and to prioritize candidate therapeutics. We investigate recent hypotheses on a systems biology level to explore mechanistic virus life cycle drivers, and to extract drug repurposing candidates. CoVex renders COVID-19 drug research systems-medicine-ready by giving the scientific community direct access to network medicine algorithms. It is available at <https://exbio.wzw.tum.de/covex/>.

¹Chair of Experimental Bioinformatics, TUM School of Life Sciences, Technical University of Munich, München, Germany. ²Institute of Virology, TUM School of Medicine, Technical University of Munich, München, Germany. ³LipiTUM, Chair of Experimental Bioinformatics, TUM School of Life Sciences, Technical University of Munich, München, Germany. ⁴Natural Sciences Department, Universidad Autónoma Metropolitana-Cuajimalpa (UAM-C), 05300 Mexico City, Mexico. ⁵Computational Biomedicine Lab, Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark. ⁶These authors contributed equally: Sepideh Sadegh, Julian Matschinske. ✉email: jan.baumbach@wzw.tum.de

Coronavirus Disease-2019 (COVID-19) is an infectious disease caused by SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2). It was first identified in Wuhan, China and has spread causing an ongoing pandemic¹ with globally 2.4 million confirmed cases and 167 thousand deaths as of April 20, 2020.

Our insights into SARS-CoV-2 infection mechanisms are limited and clinical therapy has largely focused on treating critical symptoms. Therefore, the current pandemic requires fast and freely accessible knowledge to accelerate the development of vaccines, treatments, and diagnostic tests. Research data have been collected in several online platforms, such as the COVID-19 Open Research Dataset and the Dimensions COVID-19 collection^{2,3}. In addition, existing databases that collect virus information have responded by integrating new SARS-CoV-2 research^{4,5}.

As vaccine and drug development may take years, drug repurposing is a potent approach that offers new therapeutic options through the identification of alternative uses of already approved drugs⁶. These drugs have previously undergone clinical and safety trials and, hence, accelerate drug development timelines from a decade to a few years or months. Due to the COVID-19 pandemic, numerous research groups around the world have been joining their efforts to identify drugs that can be repurposed to effectively treat COVID-19. Numerous drugs are already part of clinical trials, including Remdesivir (a less effective ebola drug), Chloroquine, Hydroxychloroquine (antimalarial drugs), Tocilizumab (rheumatoid arthritis drug), Favipiravir (influenza drug), and Kaletra (a combination of Lopinavir and Ritonavir for treating human immunodeficiency virus HIV-1)⁷.

Computational systems and network medicine approaches offer a methodological toolbox required to understand molecular virus–host–drug mechanisms and to predict novel drug targets to attack them^{8,9}. Few studies on these mechanisms in SARS-CoV-2 exist. Gordon et al.¹⁰ applied affinity purification–mass spectrometry (AP-MS) to reconstruct the SARS-CoV-2–human protein–protein interaction (PPI) network and subsequently employed a cheminformatics approach to identify potential drugs for repurposing. The data generated from that study is a major advancement in understanding SARS-CoV-2 infection. However, to identify drug candidates, the study mainly considered the direct interactors of the human proteins as putative targets and thus did not take into account the network context of the human interactome. However, viral interactions with human proteins have cascading effects in the human interactome, where key proteins necessary for the viral replication cycle are only indirectly affected. Therefore, downstream host proteins may be additional promising targets for therapeutic intervention, but require thorough data integration and mining to be identified (see Supplementary Methods for details). Figure 1 illustrates the concept of systems medicine-based drug repurposing specifically for SARS-CoV-2.

Gysi et al.¹¹ integrated the experimentally validated SARS-CoV-2 virus–host interactions with the human interactome and investigated comorbidity and differences of virus–host interactions across 56 tissues. Furthermore, network medicine analysis was applied to compile a list of drug repurposing candidates that target also indirectly affected proteins in the human interactome. However, the combined number of virus–host, host–host, and drug–target interactions goes into the millions such that purely algorithmic approaches to discovering new drug targets and drug repurposing candidates produces a large number of results, many of which lack mechanistic specificity and, hence, are not useful. Thus, to make their results accessible, Gysi et al.¹¹ worked closely together with clinical experts to narrow down the number of predicted repurposable drugs.

In order to allow for the interactive integration of expert knowledge about virus replication, immune-related biological processes, or drug mechanisms, we developed the interactive systems and network medicine platform CoVex (CoronaVirus Explorer). It integrates experimental virus–human interaction data for SARS-CoV-2 and SARS-CoV-1 with the human interactome as well as drug information to predict novel drug (target) candidates, and it offers biomedical and clinical researchers' interactive and user-friendly access to network medicine algorithms for advanced data mining and hypothesis testing. CoVex follows a human-in-the-loop paradigm and provides an intuitive visualization of virus–host interactions, drug targets, and drugs to enable researchers to examine molecular mechanisms that can be targeted using repurposed drugs. CoVex offers two main actions for which several network medicine algorithms are available: Given a list of user-selected human host proteins, viral proteins, or drugs (referred to as seeds), users can (1) search the human interactome for viable drug targets and (2) identify repurposable drug candidates. In a typical workflow, these two actions are combined, that is, starting from a selection of virus or virus-interacting proteins, users mine the interactome for suitable drug targets for which, in turn, suitable drugs are identified. Additionally, users can leverage expert knowledge by uploading a list of proteins or drugs of interest as seeds to guide the analysis. Such seeds could, for instance, be a list of differentially expressed genes (DEGs), a list of proteins related to a molecular mechanism of interest, or a set of drugs known to be effective.

The remainder of this paper is structured as follows: In the “Methods” section, we first describe the datasets and integration strategy used in CoVex. Next, we introduce the rationales of the systems and network medicine algorithms implemented in CoVex, and briefly describe the overall architecture of the platform. In the “Results” section, we show several application examples to illustrate the flexibility and typical use cases of CoVex. Finally, we will discuss opportunities and limitations in using CoVex for COVID-19 research.

CoVex opens up the systems medicine toolbox for the entire infectious disease research community by providing an easy-to-use web tool enriched with data mining algorithms for drug repurposing. This allows specialists from different fields to bring in expert knowledge to identify the most promising drug targets and drug repurposing candidates for developing effective therapies. We would like to stress that the CoVex platform can and will be adopted and extended to allow exploring other viral–host–drug interactomes, for example, with MERS (Middle East respiratory syndrome), Zika, dengue, and influenza viruses, thereby increasing preparedness for similar future events.

Results

The CoVex platform. The main result is the CoVex platform itself, which renders drug repurposing research systems-medicine-ready. In the following, we first describe how the platform's user interface (Fig. 2) provides the full feature spectrum of CoVex to clinicians and scientists. Afterwards, we demonstrate the use of CoVex in four different application scenarios starting with four hypotheses and ending with different drug repurposing candidates, as well as a short discussion on how to prioritize them (Fig. 3).

Figure 2 shows the CoVex web interface. To find potential drugs, the “Quick Start” analysis will produce a multi-Steiner tree, which considers all viral proteins as seeds and adds a small number of host proteins to connect them. Subsequently, drugs directly targeting these proteins are selected via closeness centrality. After the computation has finished, a click on the corresponding task opens the analysis results, consisting of a table

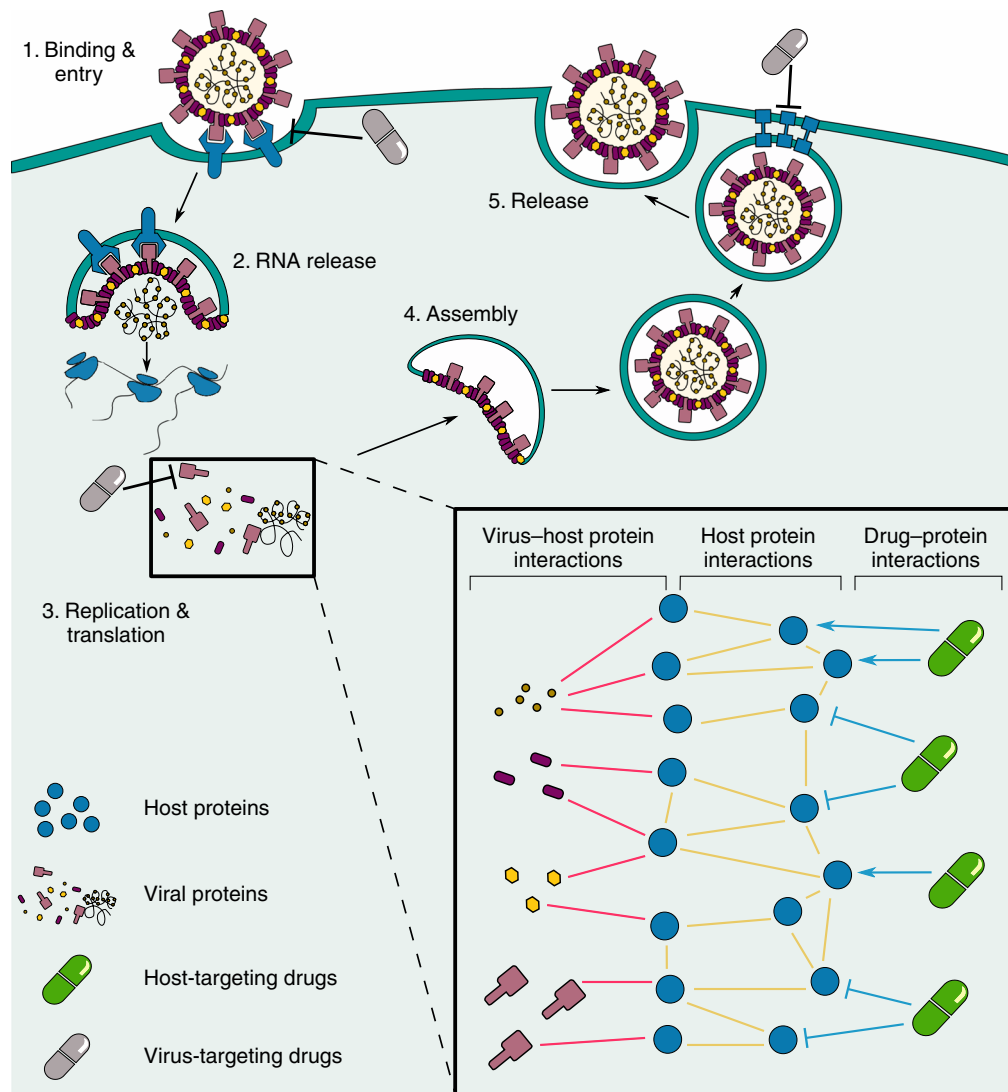


Fig. 1 The SARS-CoV-2 life cycle and the CoVex systems medicine approach of drug repurposing. Most antiviral drugs (gray drugs) target the virus proteins or their direct host interactor proteins to inhibit different stages of the viral life cycle. Our rationale, however, is that viral interactions with human host proteins have a cascading effect to hijack and control key proteins necessary for the virus’ life cycle. We aim to identify repurposable drug candidates (green drugs) targeting these key host modulators to interfere with virus replication and disease progression following infection. Besides an increased antiviral drug repertoire, targeting host proteins would make it more difficult for the virus (population) to develop resistance mutations.

view of drugs and proteins, a visualization of the protein–protein and drug–protein interactions, and a list of parameters used for the analysis. In the “Simple Analysis” panel, users can select seed proteins manually and search for drugs targeting them. In the “Advanced Analysis” panel, users can choose from a list of network medicine algorithms (see “Methods” and Supplementary Methods for details) to discover drug targets or drug repurposing candidates. Users can either select proteins from the view, upload a custom list of proteins or drugbank ids, or select proteins expressed in a given tissue. An enrichment analysis of the identified drug target proteins may be performed with g:Profiler¹².

Application scenarios. The utility of CoVex and its integrated systems medicine approaches is outlined in the following four scenarios. More details on each can be found in the Supplementary Notes.

Scenario a: Starting from a selection of viral proteins, we use the PPI network to identify the biological mechanism or pathway utilized by the virus. As an example, we consider the viral

proteins E, M, and Spike, which constitute the external structure of the virus and thus mediate entry into the host cells during the infection process^{13,14}. We select the interactors of these viral proteins reported for SARS-CoV-2 and use the multi-Steiner tree algorithm to uncover the biological pathway involved. The resulting network (Fig. 4) yields 26 new potential drug targets, including the bradykinin receptor B1 (BDKRB1). Subsequently, we use closeness centrality to find drugs affecting this pathway. Notably, we identify six relevant drugs that target BDKRB1: Ramipril, Captopril, Perindopril, and Enalaprilat (approved), which belong to the angiotensin-converting enzyme (ACE) inhibitor class¹⁵; Icatibant, an antagonist of the bradykinin receptor B2¹⁶; and bradykinin, a non-approved drug that is degraded by the ACE¹⁷. Furthermore, to understand the relationship between BDKRB1 and two proteins known to participate in the entry of the virus (angiotensin-converting enzyme 2 (ACE2) and transmembrane protease serine 2)¹⁸, we use the “custom proteins” option available in CoVex. We found that kininogen 1 and angiotensin proteins connect BDKRB1 with ACE2. These four proteins are functionally related through the

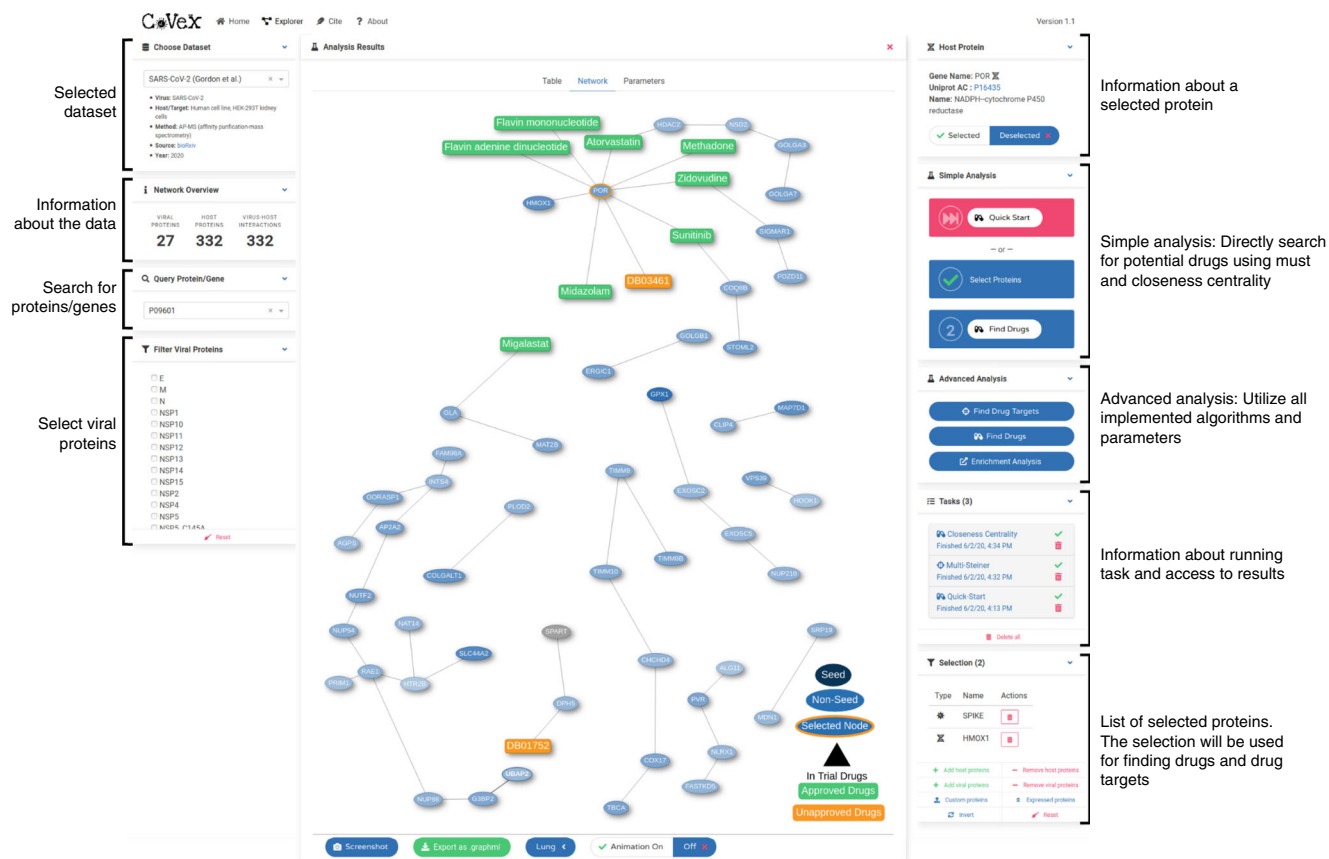


Fig. 2 The CoVex online platform. The network view (middle) shows drug candidates (green nodes) that were found using closeness centrality on a set of proteins (blue nodes), which resulted from a multi-Steiner tree computation with all viral proteins as seeds (not shown here). Therefore, drugs targeting these seeds might be able to interrupt the viral life cycle progression. Here we colored nodes based on lung-tissue-specific median gene expression according to GTEx.

renin-angiotensin system, which is targeted by ACE inhibitors (www.wikipathways.org/instance/WP554). In summary, CoVex identifies the protein BDKRB1, which appears to play a role in SARS-CoV-2 host cell entry and can be targeted by several ACE inhibitors widely used in clinical trials to treat COVID-19. It should be noted that the ACE2 protein is not present in the set of seeds used to start the analysis. Nevertheless, CoVex is capable of identifying the pathway and new protein targets functionally related to ACE2 (Fig. 4).

Scenario b: Starting from both viral proteins and a list of proteins of interest, we can use CoVex to identify a connecting pathway or biological mechanisms that can be targeted by drugs. In this scenario, we are specifically interested in viral proteins that suppress host immunity and the corresponding host immune response pathways. First, we select the viral proteins ORF7a and ORF3a, which are potentially involved in innate immune response and apoptosis as discussed by Gordon et al.¹⁰. Next, we compile a list of proteins of interest based on the DEGs from the study by Blanco-Melo et al.¹⁹ lung epithelial cells were infected with the SARS-CoV-2 virus, leading to altered expression of immunity-related genes to combat the viral infection. We consider DEGs known to be associated with the host pathway involving infection with the herpes simplex virus, another viral pathogen. These genes include *IFIH1*, *OAS1*, *STAT1*, *DDX58*, *OAS2*, *OAS3*, *IRF7*, *EIF2AK2*, *IFIT1*, and *IRF9*. The selected viral proteins and DEGs (converted to Uniprot ids) were used as seeds for the multi-Steiner tree algorithm to extract a potential immune-related mechanism. As expected, the resulting

subnetwork reveals that the viral proteins are close to the DEGs in the host PPI network. Closeness centrality analysis assigned a high rank to Tofacitinib and Ruxolitinib, which are currently being assessed in clinical trials. Tofacitinib and Ruxolitinib exert immunomodulatory effects as Janus kinase inhibitors^{20,21}. Thus, administration with these drugs may mitigate immune-mediated lung injury and reduce functional deterioration caused by an overamplified host inflammatory response. This could be especially important in later stages of the disease to prevent an overreaction of the body's immune system and, hence, may further prevent the need for mechanical ventilation in patients suffering from severe COVID-19. Other drugs that target this subnetwork include Masitinib, Erlotinib, and Sorafenib, which could be further examined in downstream analyses. In a similar manner, users may provide a custom list of proteins as seeds to hunt for drugs that can target a putative mechanism of interest.

Scenario c: Starting with a set of drugs of interest, we can follow a top-down approach to extract potential host mechanisms and additional drugs targeting the proteins participating in these mechanisms. As an example, we identify 69 drugs currently in clinical trials for COVID-19 and group them based on their Anatomical Therapeutic Chemical classification (Supplementary Table 5)²². We focus on drugs from the immunostimulants class (L03) and their target proteins as starting seeds. We further select the interactors of the immune-related viral proteins ORF9B, ORF6, ORF3B, and ORF3A¹⁰ as end-point seeds. By applying the multi-Steiner tree algorithm, we discover pathways of interacting proteins that connect the selected drugs (and their target

we investigate the pathways connecting these viral proteins with the two effective drugs Chloroquine and Favipiravir. To this end, we select two known heme binding host proteins as seeds: cytochrome b5 reductase, which interacts with the viral protein NSP7, and the viral ORF3a, which binds to heme oxygenase 1. Using KeyPathwayMiner for drug target discovery followed by closeness centrality for drug discovery, we identify methylene blue in addition to Chloroquine and Deferoxamine, which are both in COVID-19 clinical trials^{28,29}. Notably, methylene blue is approved by the Food and Drug Administration for the treatment of methemoglobinemia, which fits the investigated hypothesis (reduced oxygen-carrying capacity). Also, Deferoxamine is widely used therapeutically as a chelator of ferric ions in disorders of iron overload³⁰. However, note that the available scientific evidence for a methemoglobinemia or ferric ion imbalance caused by SARS-CoV-2 is very limited (see Supplementary Notes) and that we use this hypothesis solely to illustrate the potential of CoVex' network medicine investigation and hypothesis testing capabilities.

Discussion

COVID-19 is a threat to our health and our social life, as well as to our healthcare and economic systems around the globe. Since the development of safe and effective vaccines is a time-consuming process, the only alternative to mitigate the damage by the SARS-CoV-2 pandemic is to quickly identify agents for the treatment and control of COVID-19 symptoms. Much attention in biomedical and clinical research is, thus, given to the task of identifying therapeutically exploitable drugs. A particular interest lies in drug repurposing, since already approved drugs can go through shortened clinical trials within months rather than years. While a number of promising drug repurposing candidates are currently being tested, the discovery of such candidates is still unstandardized and mostly unstructured. Systems and network medicine offer alternative approaches, where the process of drug target discovery is driven by computational data mining methods utilizing molecular interaction networks. As recently demonstrated by Gysi et al.¹¹ for SARS-CoV-2, this data-driven process can produce a list of promising drug candidates targeting host proteins in close proximity and mechanistically related to virus-interacting proteins¹¹. Here, we seek to make this network medicine approach widely available to the community.

With CoVex, we present an interactive and user-friendly web platform that integrates published data of SARS-CoV-1 as well as recent data about virus–host interactions in SARS-CoV-2¹⁰ with the human interactome and several drug–target interaction databases. CoVex allows users to mine the integrated virus–host–drug interactome for putative drug targets and drug repurposing candidates with only a few mouse clicks. Through features such as interactive seed protein selection, filtering, and upload of own lists of proteins or drugs of interest, CoVex covers diverse application scenarios ranging from data-driven, hypothesis-free drug target discovery to expert-guided analyses with a clear underlying hypothesis about virus biology. To address the diversity of research questions adequately, CoVex implements several state-of-the-art graph analysis methods. These were specifically tailored to be employed in a network medicine context and include a weighted version of TrustRank as well as a multi-Steiner tree method (Supplementary Material).

While CoVex is a powerful tool for SARS-CoV-1 and -2 research, results uncovered with our platform have to be considered with caution. We stress that CoVex can only suggest putative drug candidates for further investigation and that those candidates are not guaranteed to have an antiviral effect. While the suggested drugs target proteins involved in a putatively important mechanism for the virus, the actual effect of the drug

has to be verified through follow-up investigations. The inhibition of a cofactor that prevents the virus from manipulating host proteins, for example, could even have a proviral effect. After validating the target for the suggested drug through appropriate genetic or chemical approaches, the drug candidate, hence, still needs to be properly vetted by clinical experts and tested following established procedures and clinical trials. Current data about virus–host interactions in SARS-CoV-2 is still preliminary and incomplete. For instance, important proteins such as the ACE2 receptor, a known entry point for the virus¹⁸, is missing in the SARS-CoV-2 dataset by Gordon et al.¹⁰. Moreover, we included only drugs that are reported in databases about clinical trials or in the literature if they have a valid entry in DrugBank, possibly excluding some of the drugs currently being investigated. Further, we do not differentiate between different sources of drug–target interactions. The strength of experimental evidence may vary depending on the experimental assay that was used or the type of annotation from the source database, for example, clinical and variant annotations from PharmGKB, which can be interpreted as indirect drug–protein associations. It should also be noted that we do not list drugs that target viral proteins directly, as the goal of CoVex is to unravel novel drug targets further downstream in the human interactome.

We acknowledge that the choice of algorithm and its associated parameters is nontrivial, forcing users to engage in time-consuming explorative analysis. To make this easier, we allow users to queue multiple tasks, which are executed in parallel. As our experience with this platform grows, we also plan to develop guidelines that allow users to choose an appropriate method for a particular research question. We further plan to integrate new data about virus–host interactions and ongoing clinical trials in corona viruses as it becomes available.

In summary, we have presented CoVex, a web-based platform for the interactive exploration and network-based analysis of virus–host interactions, aimed towards drug repurposing for the treatment of COVID-19. CoVex can be easily updated to accommodate the fast-paced data generation in the battle against the global pandemic. CoVex is expected to speed up the discovery of potential therapeutics for COVID-19. For the future, we also plan to extend the CoVex network medicine platform to other viruses in which new drug targets and drug repurposing candidates are urgently sought, including MERS, Zika, influenza, and dengue. We will also add features for the integration of additional molecular data, such as gene expression. Until then users can work with the “add custom protein” functionality of CoVex, allowing them to utilize and filter by any set of genes, including those derived by gene expression pattern analyses.

Methods

Data integration. We integrated virus–host interaction data from several sources. We obtained SARS-CoV-2 AP-MS data reported by Gordon et al.¹⁰, containing 332 high-confidence virus–host interactions for 27 SARS-CoV-2 proteins¹⁰, as well as SARS-CoV-1 interactions from VirHostNet⁴ (24 interactions), and Pfeifferle et al.³¹ (113 interactions existing in our interactome). Human PPIs were obtained from the integrated interactions database³² filtered based on experimental validation. The resulting interactome consists of 17,666 proteins connected via 329,215 interactions. Drug–target associations were obtained from ChEMBL (2020-03)³³, DrugBank (v. 5.1.5)²⁵, DrugCentral (2018-08-26)³⁴, Target Therapeutic Database (2019-07-14)³⁵, Guide To Pharmacology (2020-01; only approved drugs)³⁶, PharmGKB (downloaded 2020-04)³⁷, and BindingDB (2019-08-12)³⁸. Where applicable, we considered drugs that have binding affinity values (EC_{50} , IC_{50} , K_d , and K_i) $< 10 \mu M$ ^{39,40}. Only drugs that were mappable to DrugBank IDs and targeting host proteins were included in the network. Drugs currently undergoing clinical trials and mappable to DrugBank IDs (as of April 4, 2020) for the treatment of COVID-19 were collected from ClinicalTrials.gov (www.ClinicalTrials.gov)⁴¹, the EU Clinical Trials Register (www.clinicaltrialsregister.eu), and the International Clinical Trials Registry Platform (www.who.int/ictrp/). In total, we have 6861 drugs (67 in clinical trials) and 52,860 drug–target associations integrated in our network. We further downloaded per-tissue median gene expression levels from the GTEx

data portal (Release V8, dbGaP Accession phs000424.v8.p2, downloaded 2020-05-30) to allow for tissue-specific filtering and visualization of gene expression values. Note that we rely on integrating published data and, thus, on their corresponding quality.

Systems medicine algorithms for drug repurposing prediction. The general idea of CoVex is to provide researchers and clinicians with a tool to visually explore druggable molecular mechanisms that drive the interactions between virus and host. To this end, the integrated virus–human–drug interactions form molecular networks that are modeled as graphs with nodes as proteins or drugs, and edges referring to interactions between them. The goal of CoVex is to explore this network while allowing for the exploitation of expert knowledge. Starting with a selected set of (usually) hypothesis-driven seeds (virus proteins, human proteins, or drugs), the goal is to first identify subnetworks connecting these seeds and, subsequently, to identify drug repurposing candidates associated with these mechanisms. A vast number of methods have been reported in the literature for identifying subnetworks⁴². In CoVex, we have integrated several algorithms (including a dedicated multi-Steiner tree algorithm) with different underlying paradigms to provide specific exploration options to various particular medical, therapeutic, and research questions and hypotheses. CoVex, thus, allows users to choose among the following approaches in the “advanced analysis” procedures.

Degree centrality is the simplest conceivable centrality measure and ranks proteins or drugs interacting with the seeds by their node degree, that is, the number of interactions. Thus, this algorithm yields subnetworks in which seed-connected proteins and/or drugs are preferentially selected if they interact with many other proteins in the network. The only user-selected parameter is the result size, that is, how many of the top-ranked proteins or drugs are included. Notably, centrality measures in CoVex can be used for detecting drug targets and for identifying promising drugs.

Closeness centrality is a node centrality measure that ranks the nodes in a network based on the lengths of their shortest paths to all other nodes in the network. The rationale behind this algorithm is to preferentially select proteins and/or drugs that are a short distance from all other proteins in the network and are thus of central importance. In CoVex, we use a modified version suggested by Kacprowski et al.⁴³, where only the shortest paths to a set of selected seed nodes are considered. The only algorithm-specific, user-selected parameter is the result size.

Betweenness centrality is another node centrality measure that ranks the nodes in a network based on how many shortest paths pass through them. In CoVex, we use a modified version suggested by Kacprowski et al.⁴³, which only considers shortest paths between pairs of seed nodes. Hence, nodes receive a high score if they are on many shortest paths between the seeds. Since drugs are not contained in any shortest paths in our integrated interactome (see Fig. 1), betweenness centrality can be used only to find drug targets. The only algorithm-specific, user-selected parameter is the result size.

Guney et al.⁴⁴ introduced the network proximity between a drug and a set of seed nodes as the average minimum distance from the drug’s targets to all of the seeds. The algorithm computes empirical *z*-scores by comparing the obtained proximity score to a background distribution obtained by randomly sampling sets of seed nodes and drug targets. In CoVex, network proximity can be employed to find drugs, given a set of host proteins of interest. The user can specify the result size, as well as the number of randomly sampled instances used for computing the background distribution.

TrustRank is conceptually similar to closeness centrality but additionally considers the importance of the seed nodes themselves. In other words, TrustRank ranks nodes in a network based on how well they are connected to a (trusted) set of seed nodes⁴⁵. It is a variant of Google’s PageRank algorithm, where “trust” is iteratively propagated from seed nodes to neighboring nodes using the network structure. The node centralities are initialized by assigning uniform probabilities to all seeds and zero probabilities to all non-seed nodes. In CoVex, the TrustRank algorithm can be run starting from a user-defined set of (trusted) seed proteins to obtain a ranked list of proteins in the PPI network that could be prioritized as putative drug targets. Similarly, TrustRank can be executed on the joint protein–drug interactome to identify drug repurposing candidates. User-selected parameters include the result size and the damping factor (range 0–1), which controls how fast “trust” is propagated through the network. A small damping factor results in a conservative behavior of the algorithm (nodes close to the seeds receive much higher scores than distant ones), while a large damping factor makes its behavior more explorative.

The Steiner tree problem is a classical combinatorial optimization problem. It aims at finding a subgraph of minimum cost connecting a given set of seed nodes. For CoVex, we have developed a weighted multi-Steiner tree method that computes approximate weighted multiple Steiner trees and connects them to one subnetwork. The user can select the set of proteins of interest and extract subnetwork(s) that connect the selected seed proteins as candidate mechanism(s) involved in COVID-19 progression. In this mechanistic subnetwork(s), we can then extract essential proteins and, thus, the most promising drug targets and repurposable drugs for COVID-19. User-selected parameters include the number of Steiner trees to be merged as well as the tolerance towards accepting more expensive subnetworks (for speeding up the approximation algorithm; for details see Supplementary Methods).

KeyPathwayMiner is a network enrichment tool that identifies condition-specific subnetworks (key pathways)⁴⁶. In CoVex, we utilize the KeyPathwayMiner web service to extract a maximally connected subnetwork starting from a user-defined set of proteins of interest (seeds). The only user-selected parameter is *K*, which represents the number of permitted exception nodes, that is, proteins that were not part of the seed proteins but serve to connect them. Since these proteins act as bridges, these may represent key proteins participating in the dysregulated subnetwork even though they are not directly targeted by the virus and are therefore promising candidates for intervention. In its current implementation, exception nodes will only be added if they indeed possess a bridging characteristic and will not be shown otherwise.

Irrespective of the network analysis method used, the extracted solutions have a higher intrinsic probability to contain high-degree nodes (hubs), that is, proteins that have a large number of interactions. While these proteins are key players in the human interactome, they are not necessarily suitable drug targets as perturbing them might lead to severe unintended side effects. Since it is more likely that hub proteins are involved in several mechanisms and are not specific to the mechanism of the disease under study, users can also perform the analysis with the hub penalty, which can potentially favor more specific mechanisms related to COVID-19. To mitigate this bias, users can either select an upper bound to filter out high-degree nodes or, alternatively, penalize high-degree nodes by incorporating the degree of neighboring nodes as edge weights in the optimization. For the latter, values between 0 and 1 can be selected, where higher values correspond to a higher penalty. Both options are available in advanced analyses for all methods except for degree centrality, because its rationale is to identify hubs, and KeyPathwayMiner, which conceptually does not allow for weighted subnetwork extraction.

All network algorithms except multi-Steiner tree and KeyPathwayMiner yield scores for the nodes contained in the returned subnetwork. In the case of degree centrality, closeness centrality, betweenness centrality, and TrustRank, these scores correspond to, respectively, the number of direct interactions with the seeds, the inverse of the mean distance to the seeds, the fraction of shortest paths between the seeds passing through the node, and the “trust” on the node at termination. In all four cases, high scores indicate that the nodes are central with respect to the seeds, but the scores do not carry any intrinsic statistical semantics. In CoVex, we hence display normalized scores for degree centrality, closeness centrality, betweenness centrality, and TrustRank, which we compute by dividing by the obtained maximum. In contrast to that, network proximity yields empirical *z*-scores, which are smaller the more promising the drugs are for the selected set of seed proteins. Since these *z*-scores directly translate into empirical *p* values, we do not normalize them.

Implementation. CoVex consists of five components: (i) Data are stored in a PostgreSQL database (v. 12.2). (ii) The backend is implemented using the Django web framework (v. 3.0.5) with Python (v. 3.6) and the Django REST framework (v. 3.11.0) to build the web API. (iii) The network algorithms (except KeyPathwayMiner) are implemented with graph-tool (v. 2.3.1)⁴⁷. (iv) Background task processing is implemented using Redis Queue (RQ, v. 1.3.0) and the in-memory database Redis (v. 3.4.1). Django enqueues the jobs and RQ processes them in the background while Redis functions as a broker between Django and RQ. (v) The frontend is implemented in Angular (v. 9.0.2) and utilizes the JavaScript libraries vis-data (v. 6.5.1) and vis-network (v. 7.4.2) for network visualization.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The authors declare that all data supporting the findings of this study are available publicly and their integration is described accordingly within the paper and its supplementary information files. Human protein–protein interactions were obtained from the Integrated Interactions Database (<http://iid.ophid.utoronto.ca/>). Virus–host interactions were downloaded from VirHostNet (<http://virhostnet.prabi.fr/>). Drug–target associations were integrated from the following databases: ChEMBL (<https://www.ebi.ac.uk/chembl/>), DrugBank (<https://www.drugbank.ca/>), DrugCentral (<http://drugcentral.org/>), Target Therapeutic Database (<http://bidd.nus.edu.sg/group/cjttd/>), Guide To Pharmacology (<https://www.guidetopharmacology.org/>), PharmGKB (<https://www.pharmgkb.org/>), and BindingDB (<https://www.bindingdb.org/bind/index.jsp>). Drugs undergoing clinical trials for COVID-19 were collected from ClinicalTrials.gov (<https://clinicaltrials.gov/>), the EU Clinical Trials Register (<https://www.clinicaltrialsregister.eu/>), and the International Clinical Trials Registry Platform (<https://www.who.int/ictrp/en/>). Tissue-specific gene expression levels were obtained from the GTEx data portal (<https://www.gtexportal.org/home/>), dbGaP Accession phs000424.v8.p2).

Code availability

CoVex is a public online platform software running on a web server. The CoVex code is available from the corresponding author upon reasonable request. The online tool is available at <https://exbio.wzw.tum.de/covex/>.

Received: 23 April 2020; Accepted: 12 June 2020;

Published online: 14 July 2020

References

- World Health Organisation. Coronavirus. <https://www.who.int/emergencies/diseases/novel-coronavirus-2019> (2020).
- Dimensions Resources. Dimensions COVID-19 publications, datasets and clinical trials. https://dimensions.figshare.com/articles/dataset/Dimensions_COVID-19_publications_datasets_and_clinical_trials/11961063 (2020).
- Semantic Scholar. COVID-19 open research dataset (CORD-19). <https://pages.semanticscholar.org/coronavirus-research>.
- Guirimand, T., Delmotte, S. & Navratil, V. VirHostNet 2.0: surfing on the web of virus/host molecular interactions data. *Nucleic Acids Res.* **43**, D583–D587 (2015).
- Guirimand, T., Delmotte, S. & Navratil, V. VirHostNet 2.0: surfing on the web of virus/host molecular interactions data. *Nucleic Acids Res.* **43**, D583–D587 (2014).
- Sun, P., Guo, J., Winnenburg, R. & Baumbach, J. Drug repurposing by integrated literature mining and drug–gene–disease triangulation. *Drug Discov. Today* **22**, 615–619 (2017).
- World Health Organisation. ‘Solidarity’ clinical trial for COVID-19 treatments. <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/global-research-on-novel-coronavirus-2019-ncov/solidarity-clinical-trial-for-covid-19-treatments> (2020).
- Casas, A. I. et al. From single drug targets to synergistic network pharmacology in ischemic stroke. *Proc. Natl. Acad. Sci. USA* **116**, 7129–7136 (2019).
- Baumbach, J. & Schmidt, H. The end of medicine as we know it: introduction to the new journal, systems medicine. *Network Syst. Med.* **1**, 1–2 (2018).
- Gordon, D. E., Jang, G. M., Bouhaddou, M., Xu, J. & Obernier, K. A SARS-CoV-2-human protein–protein interaction map reveals drug targets and potential drug-repurposing. *BioRxiv* <https://doi.org/10.1101/2020.04.15.341586> (2020).
- Gysi, D. M. et al. Network medicine framework for identifying drug repurposing opportunities for COVID-19. Preprint at arXiv:2004.07229v1 [q-bio.MN] (2020).
- Raudvere, U. et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* **47**, W191–W198 (2019).
- de Haan, C. A. M. & Rottier, P. J. M. in *Advances in Virus Research*, Vol. 64, 165–230 (Academic Press, 2005).
- Walls, A. C. et al. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell* **181**, 281–292.e6 (2020).
- Piepho, R. W. Overview of the angiotensin-converting-enzyme inhibitors. *Am. J. Health Syst. Pharm.* **57**(Suppl. 1), S3–S7 (2000).
- HOE 140, JE 049, JE049. Icatibant. *Drugs R D* **5**, 343–348 (2004).
- Kuoppala, A., Lindstedt, K. A., Saarinen, J., Kovanen, P. T. & Kokkonen, J. O. Inactivation of bradykinin by angiotensin-converting enzyme and by carboxypeptidase N in human plasma. *Am. J. Physiol. Heart Circ. Physiol.* **278**, H1069–H1074 (2000).
- Hoffmann, M. et al. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *Cell* <https://doi.org/10.1016/j.cell.2020.02.052> (2020).
- Blanco-Melo, D. et al. SARS-CoV-2 launches a unique transcriptional signature from in vitro, ex vivo, and in vivo systems. *bioRxiv* <https://doi.org/10.1101/2020.03.24.004655> (2020).
- Elli, E. M., Baratè, C., Mendicino, F., Palandri, F. & Palumbo, G. A. Mechanisms underlying the anti-inflammatory and immunosuppressive activity of Ruxolitinib. *Front. Oncol.* **9**, 1186 (2019).
- van Vollenhoven, R. et al. Evaluation of the short-, mid-, and long-term effects of Tofacitinib on lymphocytes in patients with rheumatoid. *Arthritis Rheumatol.* **71**, 685–695 (2019).
- WHOCC. WHOCC-ATC/DDD Index. WHOCC https://www.whooc.no/atc_ddd_index/ (2016).
- National Library of Medicine. Cytokine signaling in immune system. *Reactome* <https://reactome.org/content/detail/R-HSA-1280215> (2020).
- Uhlen, M. et al. A pathology atlas of the human cancer transcriptome. *Science* **357**, <https://doi.org/10.1126/science.aan2507> (2017).
- Wishart, D. S. et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* **46**, D1074–D1082 (2018).
- Abrahams, L. Covid-19: acquired acute porphyria hypothesis. Preprint at <https://doi.org/10.31219/osf.io/4wkyf> (2020).
- Wenzhong, L. & Hualan, L. COVID-19: attacks the 1-beta chain of hemoglobin and captures the porphyrin to inhibit human heme. *Metabolism* <https://doi.org/10.26434/chemrxiv.11938173.v7> (2020).
- Vincent, M. J. et al. Chloroquine is a potent inhibitor of SARS coronavirus infection and spread. *Virology* **2**, 69 (2005).
- Wang, M. et al. Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-nCoV) in vitro. *Cell Res.* **30**, 269–271 (2020).
- Lederman, H. M., Cohen, A., Lee, J. W., Freedman, M. H. & Gelfand, E. W. Deferoxamine: a reversible S-phase inhibitor of human lymphocyte proliferation. *Blood* **64**, 748–753 (1984).
- Pfefferle, S. et al. The SARS-coronavirus–host interactome: identification of cyclophilins as target for pan-coronavirus inhibitors. *PLoS Pathog.* **7**, e1002331 (2011).
- Kotlyar, M., Pastrello, C., Malik, Z. & Jurisica, I. IID 2018 update: context-specific physical protein–protein interactions in human, model organisms and domesticated species. *Nucleic Acids Res.* **47**, D581–D589 (2019).
- Mendez, D. et al. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* **47**, D930–D940 (2019).
- Ursu, O. et al. DrugCentral 2018: an update. *Nucleic Acids Res.* **47**, D963–D970 (2019).
- Wang, Y. et al. Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics. *Nucleic Acids Res.* **48**, D1031–D1041 (2020).
- Armstrong, J. F. et al. The IUPHAR/BPS guide to pharmacology in 2020: extending immunopharmacology content and introducing the IUPHAR/MMV guide to malaria pharmacology. *Nucleic Acids Res.* **48**, D1006–D1021 (2020).
- Barbarino, J. M., Whirl-Carrillo, M., Altman, R. B. & Klein, T. E. PharmGKB: a worldwide resource for pharmacogenomic information. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **10**, e1417 (2018).
- Gilson, M. K. et al. BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.* **44**, D1045–D1053 (2016).
- Talevi, A. Multi-target pharmacology: possibilities and limitations of the ‘skeleton key approach’ from a medicinal chemist perspective. *Front. Pharmacol.* **6**, 673 (2015).
- Zhang, S., Zhao, H. & John, R. Development of a quantitative relationship between inhibition percentage and both incubation time and inhibitor concentration for inhibition biosensors—theoretical and practical considerations. *Biosens. Bioelectron.* **16**, 1119–1126 (2001).
- Charatan, F. US launches new clinical trials database. *BMJ* **320**, 668 (2000).
- Batra, R. et al. On the performance of de novo pathway enrichment. *NPJ Syst. Biol. Appl.* **3**, 6 (2017).
- Kacprowski, T., Doncheva, N. T. & Albrecht, M. NetworkPrioritizer: a versatile tool for network-based prioritization of candidate disease genes or other molecules. *Bioinformatics* **29**, 1471–1473 (2013).
- Guney, E., Menche, J., Vidal, M. & Barabási, A.-L. Network-based in silico drug efficacy screening. *Nat. Commun.* **7**, 10331 (2016).
- Gyöngyi, Z., Garcia-Molina, H. & Pedersen, J. Combating web spam with TrustRank. In *Proceedings 2004 VLDB Conference* (eds Nascimento, M. A. et al.) 576–587 (Morgan Kaufmann, 2004).
- Alcaraz, N. et al. Robust de novo pathway enrichment with KeyPathwayMiner 5. *F1000Res* **5**, 1531 (2016).
- Peixoto, T. P. The graph-tool python library. *Figshare* <https://doi.org/10.6084/m9.figshare.1164194> (2014).

Acknowledgements

T.K. and S.S. are grateful for financial support from H2020 project RepoTrial (no. 777111). J.M., J.S., N.K.W., and R.N. received funding from H2020 project FeatureCloud (no. 826078). J.B., T.K., and M.L. are grateful for financial support from BMBF grant Sys_CARE (no. 01ZX1908A) of the Federal German Ministry of Research and Education. J.B.’s BMBF grant SyMBoD (no. 01ZX1910D) also financed parts of this project. M.O. is funded by the Collaborative Research Centre SFB924 of the German Research Foundation. J.B. was partially funded by his VILLUM Young Investigator Grant no. 13154. Contributions by J.K.P. and T.D.R. are funded by the Bavarian State Ministry of Science and the Arts in the framework of the Center Digitization.Bavaria (ZD.B, grant LipiTUM). M.S.-A. is grateful for a Ph.D. fellowship funding from CONACYT (CVU659273) and the German Academic Exchange Service, DAAD (ref. 91693321).

Author contributions

S.S., J.M., J.B., M.L., T.K., J.K.P., A.P., and A.S. conceived and designed the study. S.S. and J.M. were in charge of overall direction, planning, and supervision. S.S., G.G., T.D.R., M.S.-A., and N.K.W. performed the acquisition, integration, and interpretation of data. S.S., D.B.B., M.L., and K.Y. developed and adapted the algorithms for network-based drug repurposing. J.M., R.N., M.O., and J.S. implemented the web platform. All authors provided critical feedback and helped in the interpretation of data, manuscript writing, and the improvement of the platform.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41467-020-17189-2>.

Correspondence and requests for materials should be addressed to J.B.

Peer review information *Nature Communications* thanks Alan Talevi, Marcel Müller and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permission information is available at <http://www.nature.com/reprints>

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









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

© The Author(s) 2020

A.2. Network medicine for disease module identification and drug repurposing with the NeDRex platform

Network medicine for disease module identification and drug repurposing with the NeDRex platform

Sepideh Sadegh ^{1,2,12✉}, James Skelton^{3,12}, Elisa Anastasi³, Judith Bennett ¹, David B. Blumenthal ⁴, Gihanna Galindez^{5,6}, Marisol Salgado-Albarrán^{1,7}, Olga Lazareva¹, Keith Flanagan³, Simon Cockell⁸, Cristian Nogales⁹, Ana I. Casas^{9,10}, Harald H. H. W. Schmidt ⁹, Jan Baumbach ^{2,11,13}, Anil Wipat^{3,13} & Tim Kacprowski ^{5,6,13}

Traditional drug discovery faces a severe efficacy crisis. Repurposing of registered drugs provides an alternative with lower costs and faster drug development timelines. However, the data necessary for the identification of disease modules, i.e. pathways and sub-networks describing the mechanisms of complex diseases which contain potential drug targets, are scattered across independent databases. Moreover, existing studies are limited to predictions for specific diseases or non-translational algorithmic approaches. There is an unmet need for adaptable tools allowing biomedical researchers to employ network-based drug repurposing approaches for their individual use cases. We close this gap with NeDRex, an integrative and interactive platform for network-based drug repurposing and disease module discovery. NeDRex integrates ten different data sources covering genes, drugs, drug targets, disease annotations, and their relationships. NeDRex allows for constructing heterogeneous biological networks, mining them for disease modules, prioritizing drugs targeting disease mechanisms, and statistical validation. We demonstrate the utility of NeDRex in five specific use-cases.

¹Chair of Experimental Bioinformatics, TUM School of Life Sciences, Technical University of Munich, Munich, Germany. ²Chair of Computational Systems Biology, University of Hamburg, Hamburg, Germany. ³School of Computing, Newcastle University, Newcastle upon Tyne, UK. ⁴Department Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander University Erlangen-Nürnberg, Erlangen, Germany. ⁵Division Data Science in Biomedicine, Peter L. Reichertz Institute for Medical Informatics of Technische Universität Braunschweig and Hannover Medical School, Braunschweig, Germany. ⁶Braunschweig Integrated Centre of Systems Biology (BRICS), TU Braunschweig, Braunschweig, Germany. ⁷Natural Sciences Department, Universidad Autónoma Metropolitana-Cuajimalpa, Mexico City, Mexico. ⁸School of Biomedical, Nutrition and Sports Sciences, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, UK. ⁹Department of Pharmacology and Personalised Medicine, School for Mental Health and Neuroscience (MHeNs), Maastricht University, Maastricht, the Netherlands. ¹⁰Department of Neurology, University Hospital Essen, Essen, Germany. ¹¹Computational Biomedicine Lab, Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark. ¹²These authors contributed equally: Sepideh Sadegh, James Skelton. ¹³These authors jointly supervised this work: Jan Baumbach, Anil Wipat, Tim Kacprowski. ✉email: sadegh@wzw.tum.de

Between 1950 and 2010, the productivity of drug development halved approximately every 9 years¹. Although this trend has changed over the past ten years², the cost of bringing a new molecular entity to market is still estimated to be between two and three billion USD³. Contributing factors to these high costs include a plethora of already effective treatments, irreproducibility of pre-clinical research and an increase of caution amongst drug regulatory agencies¹. Consequently, there is interest in alternative approaches to finding therapeutics.

Drug repurposing, also known as drug repositioning, is the process of identifying alternative uses for existing drugs. In comparison to traditional drug development, drug repurposing offers significant advantages such as low cost, reduced risk, and faster drug development timelines. While early examples of successfully repurposed drugs have been identified through serendipitous discoveries, advances in omics technologies and the availability of massive amounts of omics data have provided opportunities for systematic in silico inference of new drug-disease relationships.

Various in silico drug repurposing strategies have been proposed, including signature-, knowledge-, network-, and machine learning-based approaches⁴. Network-based approaches are particularly attractive, because networks offer a natural representation of complex biological associations and provide a framework for incorporating multiple data types. In such networks, nodes can represent drugs, proteins, or diseases, and edges indicate drug-drug similarities, drug-target interactions, gene-disease associations, and gene-gene interactions (e.g., protein-protein interaction (PPI) networks, gene regulatory networks, signaling networks, and metabolic networks)⁵.

Moreover, previous studies have indicated that disease-associated genes are not randomly scattered throughout biological networks. Instead, they tend to be located in so-called disease modules, i.e., small subnetworks representing interconnected mechanisms that can be linked to the phenotype^{6–8}. One of the guiding paradigms of network-based drug repurposing is that diseases can be viewed as perturbations of these modules⁸. Consequently, potentially repurposable drugs can be identified in silico by carrying out the following three steps:

1. Construct a heterogeneous biological network by integrating data from multiple biomedical databases which are relevant for the given task.
2. Mine the constructed biological network to derive disease modules associated with the disease of interest.
3. Extract prioritized list of drugs whose known targets are contained in or situated in close vicinity of the extracted disease modules.

Network-based drug repurposing is a highly active field of research, which has been boosted even further with the advent of the COVID-19 pandemic. However, studies have so far been limited to presenting either non-translational algorithmic results or specific predictions limited to certain diseases. There is still an urgent need for integrated tools which allow experts from pharmacology or biomedical research fields to easily carry out all three steps of network-based drug repurposing and adapt them to the needs of their individual use cases. To the best of our knowledge, the only available tools that begin to address this need are Hetionet⁵ and CoVex⁹. However, Hetionet is static and only allows the user to browse for pre-computed results related to a fixed set of 136 diseases (algorithms are provided only as separate Python packages and are not integrated into the platform). CoVex does allow the user to interact with the system, but it is limited to COVID-19 drug repurposing.

We present the NeDRex (Network-based Drug Repurposing and exploration) platform—a generically applicable integrated

platform for network-based disease module discovery and drug repurposing. Figure 1 illustrates the overview of the platform. NeDRex is built of three main components: a knowledgebase (NeDRexDB, available at <http://neo4j.nedrex.net/> and <https://api.nedrex.net/>), a Cytoscape app (NeDRexApp, available at <https://apps.cytoscape.org/apps/nedrex>), and an API (NeDRex-API, available at <https://api.nedrex.net/>).

NeDRexDB integrates data from various biomedical databases such as OMIM¹⁰, DisGeNET¹¹, UniProt¹², NCBI gene info¹³, IID¹⁴, MONDO¹⁵, DrugBank¹⁶, Reactome¹⁷, and DrugCentral¹⁸. Integration of multiple databases enables us to construct heterogeneous networks representing distinct types of biomedical entities (e.g., diseases, genes, drugs) and the associations between them. These networks can be accessed and explored via NeDRexApp, NeDRexAPI, and the Neo4j endpoint to NeDRexDB. For more details on the different types of integrated data in NeDRexDB, see Supplementary Table 1, 2 and Supplementary Fig. 1.

NeDRexApp is a Cytoscape app¹⁹ that provides implementations of state-of-the-art network algorithms, such as Multi-Steiner Trees (MuST)⁹, TrustRank²⁰, Biclustering Constrained by Networks (BiCoN)²¹, and Disease Module Detection (DIAMOND)⁸. These functionalities are made available to the user via the RESTful API and the easy-to-use NeDRexApp. All algorithms require a list of user-selected genes (referred to as seeds) as the starting point, except for BiCoN, which uses gene expression data. Seeds can be all or a subset of the genes associated with the disease, so-called disease genes, or genes contained in disease modules. Moreover, expert knowledge can be employed for seed selection, and the results can be statistically validated by calculating the empirical *P* values (Fig. 1). NeDRex, hence, allows researchers from pharmacology and biomedicine to leverage their expert knowledge for discovering drug repurposing candidates via state-of-the-art network medicine methods. In particular, our platform can also be used to identify disease modules and possibly repurposable drugs for any newly discovered disease such as COVID-19.

The remainder of the paper is organized as follows: In the Results section, we first provide an overview of the NeDRex platform. Subsequently, we present several use cases which exemplify how to use NeDRex for disease module identification and drug repurposing. In the Discussion section, we discuss prospects and limitations of using NeDRex for drug repurposing. In the Methods section, we describe the datasets and the integration scheme used in NeDRexDB. We also introduce the logic behind the network medicine algorithms implemented in NeDRex, and briefly describe the general architecture of the platform.

Results

The NeDRex platform. The main result is the NeDRex platform itself, which provides a broad spectrum of systems medicine methods together with integrative networks of different biological entities. The platform is modular and new algorithms and databases can be easily incorporated. In addition, the NeDRexDB knowledgebase, which is accessible via the RESTful API and Neo4j endpoint, serves as a useful resource for scientists to explore the relationships between different biological entities, such as drugs, diseases, genes, proteins, and pathways. Moreover, by using NeDRexApp, users can build custom networks from the NeDRexDB knowledgebase according to their needs and further explore them via the various network medicine functionalities (the complete list of functionalities is available in the tutorial document of the app: <https://nedrex.net/tutorial>). Finally, users can also download the data from NeDRexDB and employ it for

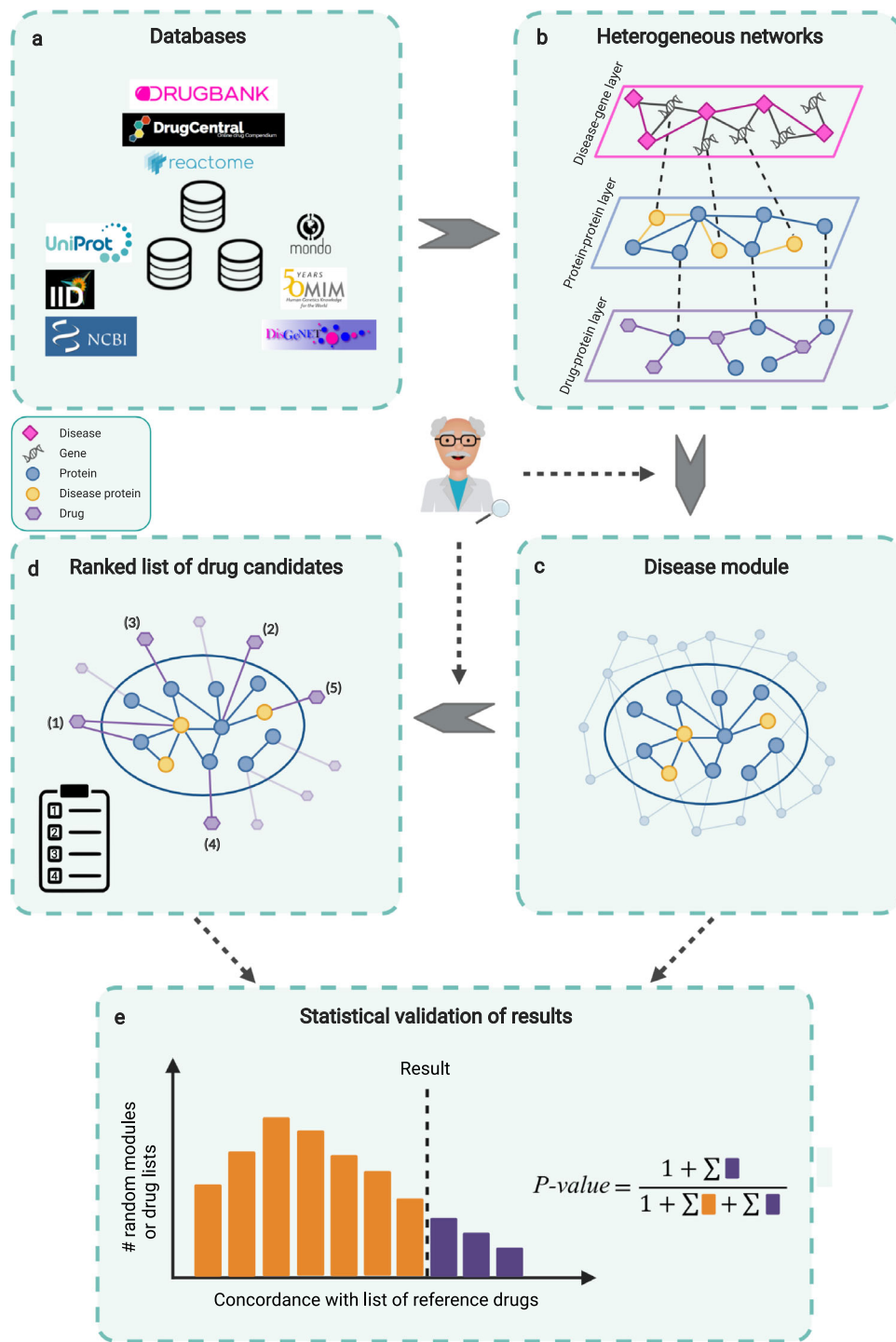


Fig. 1 Overview of the NeDRex platform. **a** Integration of various biomedical databases. **b** Construction of heterogeneous networks. **c** Disease module identification using network-based algorithms (MuST, DIAMONd, BiCoN). **d** Ranking of drugs using network-based algorithms (TrustRank, closeness centrality). Benefiting from the expert-in-the-loop paradigm, expert knowledge can be engaged at two points: (1) before the disease module identification step through selecting seeds; (2) before the drug ranking step through selecting seeds for ranking algorithms. **e** Statistical validation of the obtained disease modules and ranked lists of drugs via empirical *P* values. X-axis: Concordance of contained drugs (for drug list validation) or targeting drugs (for disease module validation) with list of reference (e.g., indicated) drugs. Created with BioRender.com.

their own drug repurposing methods. Table 1 provides an overview of the main functionalities provided by NeDRex.

The typical steps users should take in NeDRexApp to derive disease modules and pinpoint drug candidates starting with the disease(s) under study are illustrated in Fig. 2. For more information about seed selection, see Supplementary Information.

For more details on the algorithms, the selected seeds, the parameters applied for each use case and their statistical validation, see Methods and Supplementary Information (Result section), respectively. In the following, we demonstrate the applicability of NeDRex in five different use cases employing a variety of available functionalities. Detailed tutorials to reproduce

Table 1 Overview of the main functionalities of the NeDRex platform.

Functionality	Description
Integrating data from multiple biomedical databases	NeDRexDB is an integrated knowledgebase which is accessible via NeDRexAPI as well as a Neo4j endpoint.
Constructing heterogeneous biological networks from NeDRexDB	Based on users' needs, different heterogeneous networks can be constructed using NeDRexAPI or NeDRexApp.
Disease module mining	Various disease module identification algorithms can be run on NeDRexDB using NeDRexApp or NeDRexAPI, based on users' inputs.
Drug prioritization	Various drug prioritization algorithms can be run on NeDRexDB using NeDRexApp or NeDRexAPI, based on users' inputs and the results of disease module mining.
Statistical validation of results	The results of disease module identification and drug prioritization analyses can be validated with different statistical methods.
Visualization of results	Using NeDRexApp, all the obtained results are shown in network format, which can be explored further.

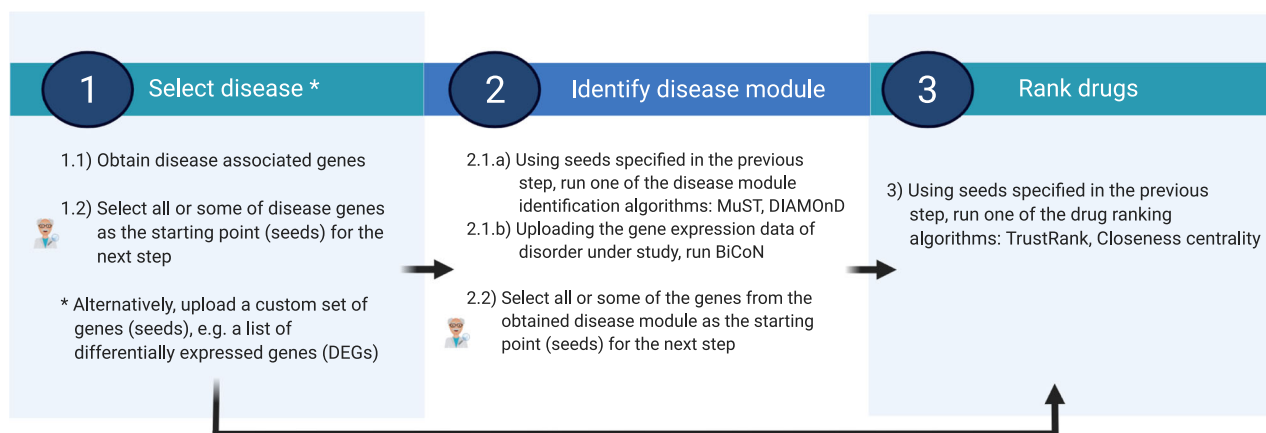


Fig. 2 Typical steps in NeDRexApp to identify disease modules and drug candidates. (Step 1) The workflow can start either with selecting the disease(s) under study and subsequently obtaining genes associated with them or uploading a custom set of genes, e.g., DEGs. (Step 2) Disease modules are derived using seeds selected in the previous step as input and employing the MuST or DIAMOnD algorithm. Alternatively, the BiCoN algorithm can be employed to return disease modules. In this case, step 1 is skipped and gene expression data should be used as input for this step. (Step 3) Drugs targeting directly or the vicinity of the seeds selected in the previous step are ranked. Step 3 can also be performed directly after step 1. Expert knowledge can be involved at seed selection points 1.2 and 2.2. Created with BioRender.com.

the use cases with NeDRexApp are available at <https://nedrex.net/tutorial>. Note that the results obtained for the use cases constitute hypotheses which have not been further experimentally validated. The main purpose of the use cases is to exemplify how to use the rich functionality available in NeDRex.

Use case 1: identification of disease pathways for ovarian cancer, using MuST. To exemplify the power of NeDRex to extract biologically meaningful pathways from starting seeds, we used the ovarian cancer (OC) associated genes from NeDRexDB (*AKT1*, *ALPK2*, *CDH1*, *CTNBN1*, *EPHB1*, *OPCML*, *PIK3CA*, *PRKN*) and constructed disease module using the MuST algorithm (Fig. 3.a). The obtained disease module contains newly identified connector genes (*ATXN1*, *HTT*, *HSP90AA1*, *PDGFRB*, *NCK1*, *OLA1* and *DKK3*) which, together with the seed nodes, participate in relevant OC pathways that could not be retrieved using the seed genes alone. In particular, genes involved in ovary-specific, hormone-related and cancer pathways are found (Fig. 3b). For instance, using the g:Profiler enrichment tool and the KEGG pathway database^{22,23}, we find that the OC module is enriched in the progesterone-mediated oocyte maturation and the Estrogen signaling pathway, which are both involved in oocyte maturation²⁴. Furthermore, we find that the ErbB signaling pathway, which is involved in cancer cell growth, proliferation, motility, and survival²⁵ is associated with the disease module. We also identified further cancer-related pathways, namely, choline

metabolism in cancer, central carbon metabolism in cancer, and EGFR tyrosine kinase inhibitor resistance^{26–28}. Finally, the examination of the connector genes identified by MuST reveals the *PDGFRB* gene, which has been reported to be deregulated in 40–80% of ovarian tumors^{29,30} and has been proposed as a therapeutic target in OC³¹.

Together, these results show that, using MuST, NeDRex was capable of identifying a disease module containing genes associated with meaningful biological pathways. Notably, although the number of seeds and the size of the disease module is small, we found ovary-specific and cancer-associated pathways, as well as genes involved in OC.

Use case 2: identification of therapeutic drugs for inflammatory bowel disease, using MuST and drug ranking algorithms.

To demonstrate the utility of the NeDRex platform to recover known and potential therapeutic drugs, we selected inflammatory bowel disease (IBD). Using the Get Disease Genes function, all the known genes associated with IBD are obtained from NeDRexDB. Running the MuST algorithm starting with this set of genes as seeds outputs a disease module containing 87 genes, which are targeted by a total of 235 drugs (empirical precision-based *P* value: 0.036). Considering the high number of drugs targeting this module, the user can prioritize the most promising candidates by using one of the drug ranking functionalities. After running the closeness centrality algorithm, three small molecules

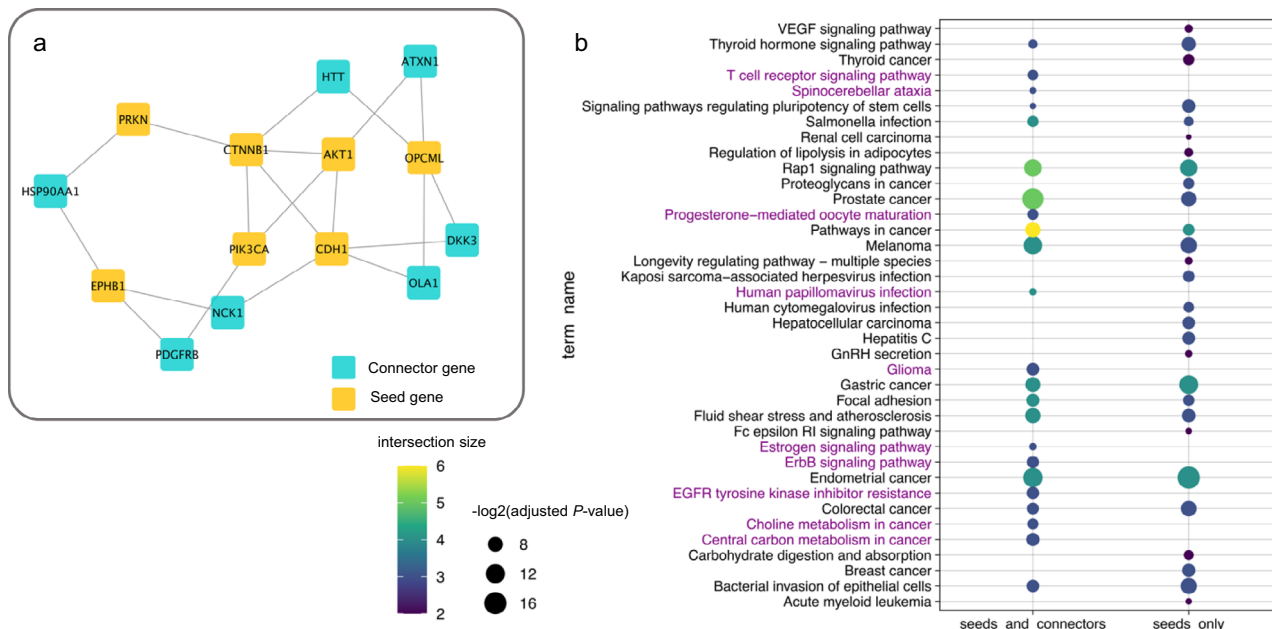


Fig. 3 Ovarian cancer disease pathway identification by MuST. **a** The OC disease module derived by MuST using NeDRexDB OC-associated genes (seeds). **b** Comparison of KEGG enriched pathways obtained with seed and connector genes vs. obtained using seed genes alone. Pathways which could only be retrieved after adding connector genes are marked in purple.

among the top-ranked drugs, namely, Fostamatinib (1), Ruxolitinib (5), and Imatinib (12) are identified, whose relevance to IBD is supported by literature evidence^{32–35}. The IBD disease module together with the 25 top-ranked drugs targeting the module is shown in Supplementary Fig. 2. Imatinib therapy has been reported to induce remission in IBD patients³². Fostamatinib was reported to alleviate IBD-induced inflammatory damage in rats³³. The JAK inhibitor Ruxolitinib has been reported to ameliorate ulcerative colitis in a mouse model³⁵.

The DCG-based empirical *P* value of the ranked list of drugs computed via closeness centrality is <0.001. The joint validation of the obtained disease module and drug list yielded a precision-based empirical *P* value of <0.001. Overall, these results provide further motivation to explore the potential of other top-ranked drugs in the treatment of IBD derived by the two algorithms using NeDRex.

Use case 3: drug target and drug identification for pulmonary embolism, using combination of DIAMOnD and TrustRank.

Next, we demonstrate how NeDRex can uncover a pulmonary embolism (PE) disease module using the DIAMOnD algorithm and subsequently recover drugs indicated for treatment of PE. Using data from NeDRexDB, twelve genes are found to be associated with PE. When selecting all of these genes as starting seeds, the DIAMOnD algorithm returns a subnetwork of 32 genes representing the underlying mechanistic pathways for PE (precision-based empirical *P* value: 0.012). A total of 283 drugs target this module. By employing the TrustRank algorithm to prioritize the drugs associated with the disease module (excluding the initial seeds), we find Bemiparin, Edoxaban, Apixaban, Dabigatran etexilate, Heparin, Rivaroxaban, Streptokinase, and Urokinase among the 50 top-ranked drugs. All of these drugs are indicated to reduce the risk of stroke and systemic embolism and are known to be used to treat PE. Furthermore, five drugs registered in ClinicalTrials.org for evaluation in treatment of PE, namely, Alteplase, Enoxaparin, Fondaparinux, Tenecteplase and Tranexamic acid are found on the top of the ranked list.

The PE disease module (excluding the initial seeds) combined with its targeting top-ranked drugs is shown in Fig. 4. Apixaban, Bemiparin, Dabigatran etexilate, Edoxaban, Enoxaparin, Fondaparinux, Heparin, and Rivaroxaban target the coagulation factor X (F10), which is not among the initial set of PE-associated genes but is found in the PE module. F10 is a key enzyme in the coagulation cascade³⁶. Alteplase, Dabigatran etexilate, Streptokinase, Tenecteplase, Tranexamic acid, and Urokinase target plasminogen (PLG), another member of the PE disease module that helps dissolving the fibrin of blood clots and behaves as a proteolytic factor³⁶. Another gene found in the PE disease module is *SERPINE1*, whose product (plasminogen activator inhibitor 1) is a protease inhibitor that is targeted by Alteplase, Tenecteplase, and Urokinase from the list of predicted drugs. This protein is essential for inhibiting fibrinolysis and is in charge of the controlled degradation of blood clots^{37,38}.

The DCG-based empirical *P* value of the ranked list of drugs computed using TrustRank is <0.001 (precision-based *P* value obtained by joint validation of module and drug list: 0.018). This use case indicates, firstly, that NeDRex is capable of extracting disease-related mechanistic pathways, which can contain possible targets for candidate drugs. Secondly, drugs which in practice are prescribed for treatment of PE or are under evaluation in clinical trials are among the top-ranked drugs obtained by the drug ranking algorithms.

Use case 4: disease module and drug identification for Huntington’s disease, using BiCoN and TrustRank.

BiCoN is an unsupervised approach that simultaneously performs patient and gene clustering such that the genes that provide the best possible clustering are also connected in a PPI network. We use BiCoN on Huntington’s disease (HD) gene expression data from GEO (accession number GSE3790^{39,40}), which contain patients with Vonsattel grades 2–4 and healthy controls (precision-based empirical *P* value of the obtained HD disease module: 0.180). Patient clusters reported by BiCoN show strong correlation with the known phenotype (average Jaccard index 0.76), providing

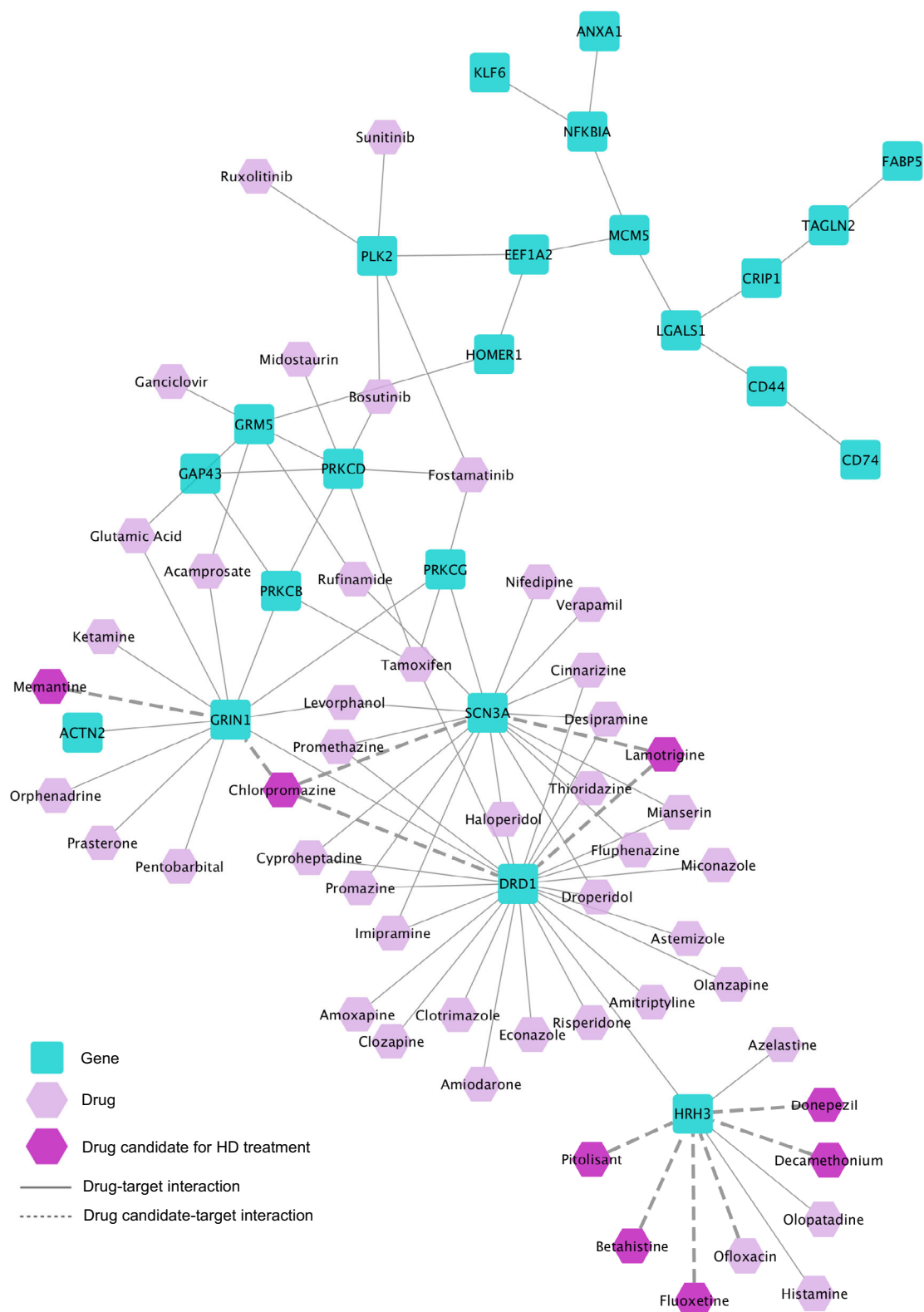


Fig. 5 The Huntington's disease module and its targeting top-ranked drugs. The HD disease module derived by BiCoN using gene expression data, together with its targeting 50 top-ranked drugs.

interest. More specifically, using Alzheimer's disease (AD) as an example, we show that we can retrieve potential treatments with an original indication for hypertension, diabetes mellitus (DM) and hyperlipidemia⁴⁹.

Hypertension as original indication - Here, we demonstrate how our platform can identify repurposable drugs directly from the genes associated with the new indication (AD) as a starting point. First, we obtain the genes associated with AD (40 genes). Then, we rank all the 240 drugs targeting this set of genes using the closeness centrality algorithm (DCG-based empirical P value: <0.001). Interestingly, this returns Telmisartan (ranked 26th). Telmisartan is a known angiotensin II receptor blocker (ARB) originally indicated to treat high blood pressure and has been tested in clinical trials to assess its efficacy for the treatment of AD⁵⁰. Studies show that drugs used to treat hypertension, including ARBs, decrease the risk and slow the progression of AD^{51,52} by reducing the amyloid- β deposition in senile plaques, the main pathological hallmark of AD. ARBs are thought to improve amyloid- β deposition through the modulation of cerebral blood flow and superoxide production⁵³. This example, hence, shows that it is possible to retrieve potentially repurposable drugs directly from the associated genes of the new indication.

Diabetes as original indication—Medications indicated for diabetes mellitus (DM) are potential treatments of AD since the glucose metabolism plays a key role in neural function^{54,55}. Several drugs have been tested in vitro, in vivo and in clinical trials, where Insulin (DB00030), Insulin Detemir, Insulin Glulisine (insulin analogs) stand out^{56–58}. These drugs interact with the insulin receptor (INSR) and are considered disease modifying drugs. Hence, we demonstrate that our platform is capable of retrieving this shared molecular mechanism and these drugs.

First, we obtain the DM-associated genes (88 genes), as well as the AD-associated genes (40 genes). The intersection of these sets consists of 2 genes: *INS* (whose encoded peptide, insulin, is a repurposed drug in AD) and *INSR* (P value = 0.017071, hypergeometric test for overlap of two disease gene sets). NeDRexDB contains 32 drugs targeting the products of these 2 genes (overlap-based empirical P value: 0.002). Notably, 27 of these drugs target INSR including repurposed drugs; such as Insulin Detemir and Insulin Glulisine. Note that, in this use case, we did not use any network algorithms to extract the drug repurposing candidates but only leveraged the data integration functionalities provided by NeDRex.

Hyperlipidemia as original indication—With this example, we show how to search for potentially repurposable drugs by retrieving drugs that indirectly target the intersection of disease modules for two diseases, namely, hyperlipidemia and AD. We use the hyperlipidemia-associated genes, since the lipid and cholesterol metabolism has been linked with progression of AD⁵⁹.

First, by using NeDRexDB, we extract the hyperlipidemia-associated genes (19 genes) and derive the disease module using DIAMOND. Similarly, we derive the AD module starting with its associated genes (40 genes). By obtaining the intersection of the two modules, we find 7 genes in common (P value of hypergeometric test = 0.023827): *A2M*, *APOE*, *APP*, *CLU*, *IGF2*, *NOS3*, and *PLAU* (precision-based empirical P value of intersection: 0.079). Notably, all of them are AD-associated genes and some are well-characterized drivers of this disease; for instance, *APP* encodes the amyloid- β peptides⁶⁰, *A2M* is a marker of neural damage⁶¹, and *APOE*, *CLU* and *NOS3* polymorphisms are risk markers of AD⁶². Importantly, *A2M*, *APP*, *CLU*, *IGF2* and *PLAU* are not among the hyperlipidemia associated genes, they are retrieved only after obtaining the disease module with DIAMOND. This demonstrates that in some cases, using only the

disease associated genes is not enough to uncover the molecular mechanisms shared between diseases and using the disease module provides a more complete landscape of the disease.

Next, to retrieve the drugs directly targeting the overlapping genes (direct drugs) or their vicinity (indirect drugs), we use closeness centrality with the option of including indirect drugs (DCG-based empirical P value of obtained ranked list of drugs: <0.001). We find Gemfibrozil among the top-ranked drugs (rank 6), which is originally indicated for the treatment of hyperlipidemia. Gemfibrozil is being tested in clinical trials (NCT02045056) and preclinical studies⁶³ give evidence of potential effectiveness of this drug for the treatment of AD. Remarkably, this drug does not directly target any of the gene products of the 7 overlapping genes, and can only be retrieved by using the indirect mode. The indirect drugs can be interpreted as drugs whose targets are closely related to the seeds; in this case, Gemfibrozil targets TTR, CYP2C8 and LPL, which interact with APOE, A2M, CLU and APP (Fig. 6), suggesting that this drug could have a positive effect by affecting several targets which altogether affect the key disease components of AD and hyperlipidemia.

Discussion

Studies in the field of drug repurposing have so far been restricted to present either non-translational algorithms or specific predictions for certain diseases. Therefore, there is an ongoing need for integrated tools which allow experts from pharmacology or biomedical research fields to easily utilize network-based drug repurposing methods and adapt them to their individual use cases.

With NeDRex, we introduce an integrated, user-friendly platform, which allows non-computer scientists and clinicians to mine different layers of a large heterogeneous biological network—the NeDRexDB knowledgebase. NeDRex provides users with a variety of network-based methods (available via NeDRexApp) to derive disease modules associated with diseases under study and prioritize drugs directly or indirectly targeting the disease modules. NeDRex also has the feature to provide prioritization for only approved drugs, which accelerates the drug development process by skipping the pre-clinical research phase and going directly into clinical trials. Benefiting from the expert-in-the-loop paradigm, researchers from biomedical sciences can leverage their domain knowledge at different points of the workflow, e.g., by filtering disease genes already provided by the platform or by using their own sets of genes as starting points for the algorithms. NeDRex hence enables researchers and clinicians to derive disease modules, explore disease-associated mechanisms, and identify drug repurposing candidates associated with these mechanisms.

We have presented five use cases which demonstrate that NeDRex can be used to mine biologically meaningful candidate disease modules as well as potentially repurposable drugs. In particular, we have shown that by using the functionalities available in NeDRex, we can identify candidate drugs that can be further explored for the treatment of inflammatory bowel disease, pulmonary embolism, Huntington's disease, and Alzheimer's disease. All results were statistically validated by empirical P values. Employing different validation methods for the use cases presented in the Results section, we computed 33 P values, 29 of which were statistically significant with significance level 0.05 (lists of all computed P values can be found in the Supplementary Information).

While the expert-in-the-loop paradigm is one of the main advantages of the NeDRex platform, it is also its most important limitation. When using NeDRex, investing domain knowledge is

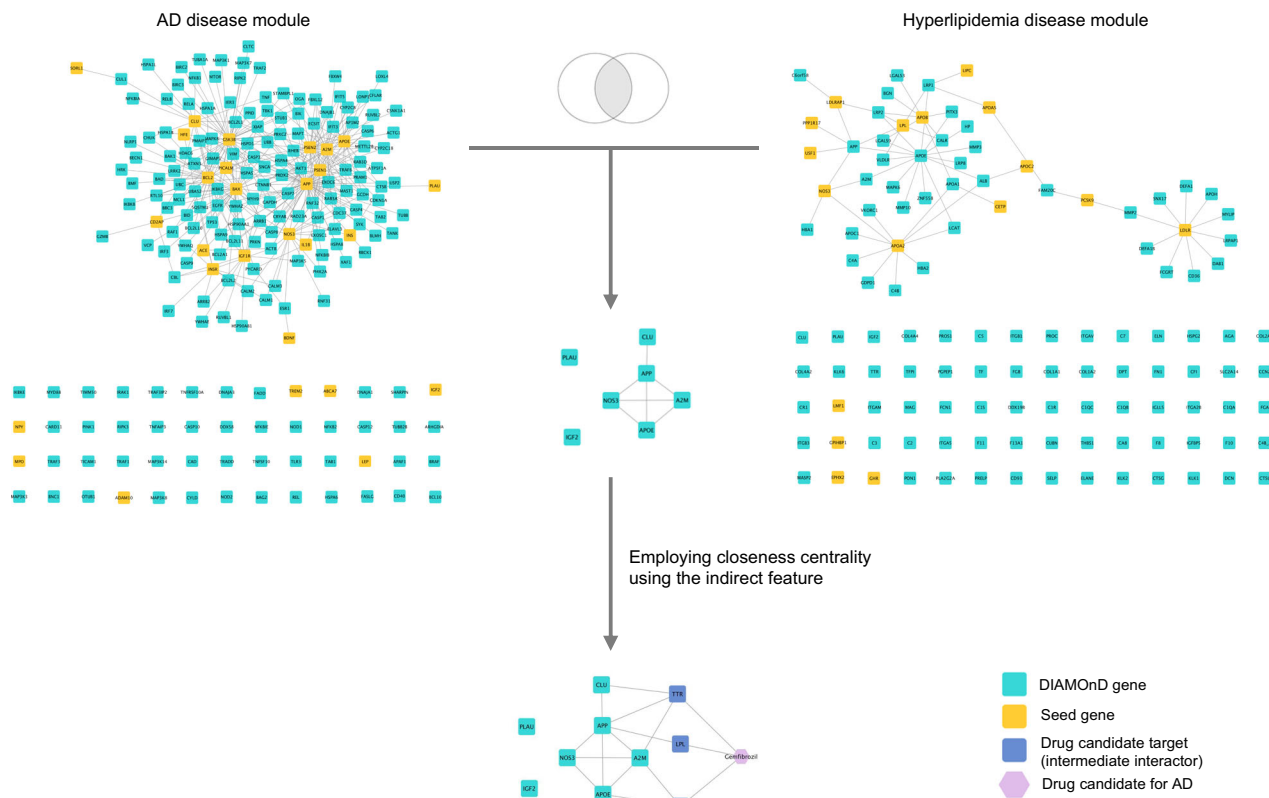


Fig. 6 Gemfibrozil indirectly targets the intersection of disease modules. The AD and hyperlipidemia disease modules (top left and top right, respectively) derived by DIAMOnD using the corresponding disease-associated genes (orange nodes). The intersection of the disease modules is shown in the middle. Gemfibrozil indirectly targets the intersection through TTR, CYP2C8, and LPL (bottom). To allow better visualisation, subsets of actual networks corresponding to the disease modules are shown here.

not an option but a requirement. If used blindly, obtaining biologically meaningful disease modules or promising drug repurposing candidates is unlikely. Importantly, also the empirical *P* values cannot replace the expert user, because they, too, are conditional on current knowledge (see “Methods” for details).

As stated above, NeDRex can only deliver putative drug candidates for further evaluation. Whereas the proposed drugs target proteins involved in potentially important disease mechanisms, the efficacy of the drug candidates needs to be verified by follow-up investigations and tested according to established rules and guidelines for clinical trials.

Finally, the integrated databases have their inherent limitations, which are reflected in our platform as well. Such limitations include false positive PPIs⁶⁴, literature bias due to under- and over-studied genes⁶⁵, and the fact that drug-protein associations available in the integrated databases do not distinguish between activation and inhibition.

For future versions of the database, we are planning to integrate disease symptoms and drug side effects data, which will allow investigation into different disease similarity and drug repositioning approaches. Regarding drug indications, previous studies (e.g., RepoDB⁶⁶) include instances of failed drugs which act as false negatives for drug indications. This has a number of advantages, such as not requiring closed-world assumptions to be made, and NeDRexDB could benefit from including similar data (e.g., from ClinicalTrials.gov). Finally, we are planning to integrate further drug repurposing databases that include tissue-level gene expression information which could help to understand why specific molecular mechanisms only lead to diseases in specific tissues.

Methods

Data integration and construction of NeDRexDB. NeDRexDB is a graph database that was constructed by integrating 10 source databases using a crowdsourcing framework. These 10 databases with their corresponding versions are shown in Supplementary Table 1. For all 10 databases, we wrote parsers to extract entities (nodes) and the relationships between entities (edges), and store them in a MongoDB instance. MongoDB was chosen as the database for two primary reasons; firstly, MongoDB has a flexible schema, which provides the freedom to readily add new characteristics to documents in the database, whilst simultaneously allowing selective enforcement of certain guarantees. Secondly, MongoDB provides a rich set of operations for querying and updating, which facilitates data integration. For more details about the data integration see Supplementary Information.

To facilitate integration, each entity in NeDRexDB was given a primaryDomainId of the form {database}.{identifier} (e.g., uniprot.P51587 for the Homo sapiens BRCA2 protein). In the cases of Proteins, Genes, and Pathways, all of the databases integrated here use UniProt, Entrez, and Reactome respectively, and so integration can be done simply on identifiers. For Drugs, DrugBank identifiers were chosen as the primary ID because DrugCentral tends to cross reference drugs to DrugBank identifiers.

Integration of diseases was more challenging, as there are no consistent identifiers used between different databases. Furthermore, mappings between disease identifiers in different databases are not complete, and many datasets do not have a hierarchy in disease concepts. Capturing a disease hierarchy in the NeDRexDB was important, as many diseases have very precise sub-typing which, for some analyses, may be too specific. We decided to use the Monarch Disease Ontology (MONDO) as the primary identifier for diseases, as the mapping between MONDO and other identifiers (e.g., the Unified Medical Language Systems (UMLS), used by DisGeNET) is more complete than others [<https://www.disgenet.org/downloads>], and includes a hierarchy.

Accessing NeDRexDB. The NeDRexDB can be accessed in two ways. The first is through a RESTful API, available at <https://api.nedrex.net/>, and the second is through a Neo4j endpoint, available at <http://neo4j.nedrex.net/>.

The routes from the API make a range of services available, including obtaining nodes and edges from NeDRexDB, ID mapping, and traversing the MONDO disease hierarchy. In addition, the API makes routes available for constructing

networks in graphml format based on users selected specifications. Graph construction is highly configurable, with options allowing filtering based on attributes (such as drug groups, IID evidence types, thresholds of gene-disease associations from DisGeNET). The documentation for the routes can be found at <https://api.nedrex.net/>. An overview of all the node and edge types available in the NeDRexDB metagraph is illustrated in Supplementary Fig. 1 and also given in Supplementary Table 2 with their corresponding numbers.

The MongoDB representation of the data was imported into a Neo4j instance, allowing users to run Cypher queries, and thus have even finer control over queries than the API allows. One major difference between the Neo4j endpoint and the API is that drugs obtained via the API are collapsed into a single Drug type by default, whereas the Neo4j instance divides these into two types, BiotechDrugs and SmallMoleculeDrugs—the abstraction used by DrugBank where drugs are sourced from.

Network-based algorithms for disease module identification and drug repurposing.

In NeDRex, we have implemented several well-established network medicine algorithms to provide various investigation options for numerous particular medical, therapeutic, and research questions. The available algorithms are detailed below. NeDRexApp allows users to select among these algorithms. Note that, although the NeDRexDB contains also predicted PPIs, only experimentally validated PPIs are considered for the networks on which the algorithms are run.

MuST—The Steiner tree problem is an optimization problem whose objective is to find a tree of minimum cost connecting the set of seeds (terminals)⁶⁷. For NeDRex we established a multi-Steiner trees method that aggregates several approximates of Steiner trees into a single subnetwork. By selecting genes associated with a disease under study as seeds, MuST extracts a connected subnetwork which potentially incorporates the genes involved in the disease pathways and mechanism. The motivation behind returning multiple trees instead of one is that the solutions to the Steiner tree problem are usually non-unique and computing several Steiner trees increases the stability of the extracted mechanism. Hub nodes, i.e., proteins having high number of interactions in the interactome, inherently have a higher chance of appearing in the extracted trees. In order to penalize the hubs and consequently extract mechanisms more specific to the disease of interest, users can conduct the MuST algorithm with the hub penalty parameter. This parameter incorporates the degree of neighboring nodes as edge weights in the optimization. In NeDRex, the MuST algorithm is implemented on the protein-protein layer of the heterogeneous network to obtain disease modules which could contain targets of putative drug repurposing candidates.

DIAMOnD—DIAMOnD⁸ identifies a candidate disease module around a set of known disease genes (seeds) by greedily adding nodes with a high connectivity significance to the module, i.e., nodes in whose neighborhoods nodes already contained in the module are significantly overrepresented. In the iterative algorithm of DIAMOnD, the connectivity significance of all direct neighbors of seeds is computed. Then, the most significantly connected node is integrated into the module, leading to expansion of the module by one node per iteration. Subsequently, the connectivity significance is recomputed w.r.t. the updated module and the process iterates until the desired module size has been reached. In contrast to MuST, DIAMOnD does not necessarily return a connected subnetwork as the disease module. In our platform, the DIAMOnD algorithm is applied to the protein-protein layer of the integrated network to derive disease modules which could incorporate targets of potential drug repurposing candidates.

BiCoN—BiCoN is a network-constrained biclustering method that is used for integrative analysis of gene expression and PPI networks²¹. BiCoN simultaneously clusters patients and genes such that genes also form a connected subnetwork in the PPI network. As an unsupervised method, BiCoN does not need a known phenotype for patients, which allows it to find entirely data-driven patients subgroups.

Closeness centrality—Closeness centrality is a node centrality measure that prioritizes the nodes in a network based on the lengths of their shortest paths to all other nodes in the network. In NeDRex, we implemented a modified version, where closeness is calculated with respect to only the selected seeds. The motivation behind this modification is to favorably select drugs that are at a close distance to the nodes in the disease module and are hence good candidates as repurposable drugs. Our implementation focuses on the combination of protein-protein and protein-drug layers of the heterogeneous network which result in a ranked list of drugs.

TrustRank—TrustRank is a modification of Google's PageRank algorithm, where the initial trust score is iteratively propagated from seed nodes to adjacent nodes using the network topology. It prioritizes nodes in a network based on how well they are connected to a (trusted) set of seed nodes²⁰. In NeDRex, it is executed on the combination of protein-protein and protein-drug layers of the heterogeneous network to obtain a ranked list of drugs that could be putative drug repurposing candidates. The damping factor parameter (range 0–1) controls the rate of trust propagation across the network. A higher damping factor returns results in a more explorative fashion.

Statistical validation. To validate the statistical significance of the lists of drugs and disease mechanisms returned by NeDRex, we have implemented three validation methods, each with two variations, based on empirical *P* values. These

validation methods allow the user to assess the statistical significance of the results obtained via different algorithms available in NeDRex, and hence make the algorithms and their results assessable and comparable w.r.t. validity and relevance. As reference, all three validation methods require a list of drugs indicated for the treatment of the disease under scrutiny. This list can either be provided by the user or be obtained directly from NeDRexDB. All empirical *P* values depend on the quality of the list of reference drugs. If this list is incomplete or contains many false positives, the *P* values might be misleading. Consequently, also the *P* values are conditional on current knowledge and therefore cannot substitute, but merely assist, the expert in the loop. The reported *P* values in the Results section and Supplementary Information are rounded to three significant digits and values smaller than 0.001 were indicated correspondingly.

a) Validation of drug lists computed by NeDRex—First, a big number of, e.g., 1000 (user parameter) ranked lists of randomly selected drugs, matching the size of the drug list predicted by NeDRex, are generated. For the predicted and each of the randomly selected drug lists, we compute the discounted cumulative gain (DCG)⁶⁸ defined as $DCG = \sum_{i=1}^n \frac{d_i}{\log_2(i+1)}$, where *n* is the length of the ranked list of drugs, $d_i = 1$ if the i^{th} drug from the sorted list of drugs is indicated for the disease of interest and $d_i = 0$ otherwise. Subsequently, an empirical *P* value is computed by counting the number of random drug lists whose DCGs exceed the DCG of the drug list predicted by NeDRex. We also implemented a simplified version (overlap-based) where, instead of the DCG, the overlap $\sum_{i=1}^n d_i$ with the reference list is used. However, unlike the DCG-based *P* values, this approach ignores whether the reference drugs are found early or late in the lists of drugs. Hence, it is recommended to be used if the user wishes to ignore the drug ranks for the statistical validation.

b) Validation of disease modules computed by NeDRex—This method takes into account the role of the disease module identification step in the NeDRex drug repurposing pipeline. We generate a number of, e.g., 1000 mock modules matching the size and the number of connected components of a disease module returned by NeDRex. We set the latter constraint to keep the topology of random modules similar to the result disease module. For the disease module computed by NeDRex as well as each mock module, we define its precision as the number of reference drugs targeting the module divided by the overall number of drugs targeting the module. We then compute an empirical *P* value by counting the number of mock modules with higher precision values than the disease module computed by NeDRex. We have also implemented a simplified approach where we do not normalize by the overall number of targeting drugs, i.e., compare intersection sizes with the reference drugs instead of precision values as defined above. If users are more interested in inspecting the number of drugs targeting a disease module, they can use the simpler version.

c) Joint validation of disease modules and drug lists computed by NeDRex—In this approach, both steps of the drug repurposing pipeline, i.e., disease module identification and drug ranking, are taken into account for the final in silico validation of drugs. Computationally, this approach is similar to the validation method for disease modules described previously. The only difference is that we now calculate the precision for the NeDRex result as the number of reference drugs contained in the drug list computed by NeDRex divided by the overall number of drugs in the list. Analogously, we use the drug lists returned by NeDRex to calculate the intersection size for the disease module computed by NeDRex. Precision values and intersection sizes for the mock modules are determined as before.

Implementation. Four modules compose the NeDRex platform: (i) NeDRexDB and its constituent metagraph. Two implementations of the NeDRexDB are used: one in Neo4j and one in MongoDB. The MongoDB version of the database is populated first, as described in the data integration section, and the MongoDB version is then exported to Neo4j. Both versions of the database are used in the API implementation, leveraging the query system advantages of both platforms. (ii) The Backend including some network-based algorithms (such as DIAMOnD, BiCoN, TrustRank and closeness centrality) is implemented with Python (v. 3.7.6). DIAMOnD was obtained from <https://github.com/dinaghiassian/DIAMOnD>, using the 22nd Sept 2020 commit (hash beginning 2437974). BiCoN was obtained from the Python Package Index (version 1.2.11). The ranking algorithms are implemented using the `graph-tool` library (v. 2.35). (iii) NeDRexAPI was constructed in Python 3 using the `fastapi` library (v. 0.61.0). (iv) NeDRexApp for Cytoscape 3 is written in Java (JDK 8). NeDRexApp serves as the primary frontend for the NeDRex platform. In addition, NeDRexApp can be used as a stand-alone app which provides access to some functions outside of the NeDRex ecosystem. For example, the MuST algorithm is implemented in both the backend as a Java command line tool and also in NeDRexApp (JDK 8) – the latter allows users to run MuST on any custom PPI network loaded into Cytoscape.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The authors declare that the NeDRexDB knowledgebase supporting the findings of this study are available via <https://api.nedrex.net/>. The construction of NeDRexDB is

described accordingly within the paper and its Supplementary Information files. The NeDRexDB knowledgebase contains information obtained from the Online Mendelian Inheritance in Man® (OMIM®) database, which has been obtained through a license from the Johns Hopkins University, which owns the copyright thereto. Use of the NeDRex dataset is governed by an End User License Agreement (available at <https://nedrex.net/about.html>), due to requirements of including OMIM as a source database.

The following databases are used in this study: IID (<http://iid.ophid.utoronto.ca/>), DrugBank (<https://go.drugbank.com/>), DrugCentral (<https://drugcentral.org/>), DisGeNET (<https://www.disgenet.org/>), OMIM (<https://omim.org/>), NCBI gene info (<https://www.ncbi.nlm.nih.gov/gene/>), UniProt (<https://www.uniprot.org/>), MONDO (<https://mondo.monarchinitiative.org/>) and Reactome (<https://reactome.org/>).

Code availability

NeDRex is a public platform built of three main components: a knowledgebase (NeDRexDB, available at <http://neo4j.nedrex.net/> and <https://api.nedrex.net/>), a Cytoscape app (NeDRexApp, available at <https://apps.cytoscape.org/apps/nedrex/>), and an API (NeDRexAPI, available at <https://api.nedrex.net/>). The NeDRexDB, NeDRexAPI, and NeDRexApp code is openly available on GitHub repositories (<https://github.com/repotrial/nedrex> and <https://github.com/repotrial/NeDRexApp>) under the terms of the GNU General Public License, Version 3.

Received: 28 July 2021; Accepted: 4 November 2021;

Published online: 25 November 2021

References

- Scannell, J. W., Blanckley, A., Boldon, H. & Warrington, B. Diagnosing the decline in pharmaceutical R&D efficiency. *Nat. Rev. Drug Discov.* **11**, 191–200 (2012).
- Ringel, M. S., Scannell, J. W., Baedeker, M. & Schulze, U. Breaking ErooM's Law. *Nat. Rev. Drug Discov.* (2020) <https://doi.org/10.1038/d41573-020-00059-3>.
- Pushpakom, S. et al. Drug repurposing: progress, challenges and recommendations. *Nat. Rev. Drug Discov.* **18**, 41–58 (2019).
- Park, K. A review of computational drug repurposing. *Transl. Clin. Pharm.* **27**, 59–63 (2019).
- Himmelstein, D. S. et al. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *Elife* **6**, e26726 (2017).
- Goh, K.-I. et al. The human disease network. *Proc. Natl Acad. Sci. USA.* **104**, 8685–8690 (2007).
- Menche, J. et al. Disease networks. Uncovering disease-disease relationships through the incomplete interactome. *Science* **347**, 1257601 (2015).
- Ghiassian, S. D., Menche, J. & Barabási, A.-L. A Disease Module Detection (DIAMOND) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. *PLoS Comput. Biol.* **11**, e1004120 (2015).
- Sadegh, S. et al. Exploring the SARS-CoV-2 virus-host-drug interactome for drug repurposing. *Nat. Commun.* **11**, 3518 (2020).
- Amberger, J. S., Bocchini, C. A., Scott, A. F. & Hamosh, A. OMIM.org: leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Res.* **47**, D1038–D1043 (2019).
- Piñero, J. et al. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.* **48**, D845–D855 (2020).
- UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).
- Maglott, D., Ostell, J., Pruitt, K. D. & Tatusova, T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.* **33**, D54–D58 (2005).
- Kotlyar, M., Pastrello, C., Malik, Z. & Jurisica, I. IID 2018 update: context-specific physical protein-protein interactions in human, model organisms and domesticated species. *Nucleic Acids Res.* **47**, D581–D589 (2019).
- Mungall, C. J. et al. The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res.* **45**, D712–D722 (2017).
- Wishart, D. S. et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* **46**, D1074–D1082 (2018).
- Jassal, B. et al. The reactome pathway knowledgebase. *Nucleic Acids Res.* **48**, D498–D503 (2019).
- Ursu, O. et al. DrugCentral 2018: an update. *Nucleic Acids Res.* **47**, D963–D970 (2019).
- Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
- Gyöngyi, Z., Garcia-Molina, H. & Pedersen, J. Combating Web Spam with TrustRank. Proceedings 2004 VLDB Conference 576–587 (2004) <https://doi.org/10.1016/b978-012088469-8.50052-8>.
- Lazareva, O. et al. BiCoN: Network-constrained biclustering of patients and omics data. *Bioinformatics* (2020) <https://doi.org/10.1093/bioinformatics/btaa1076>.
- Raudvere, U. et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* **47**, W191–W198 (2019).
- Kanehisa, M. et al. KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* **36**, D480–D484 (2008).
- Fair, T. & Lonergan, P. The role of progesterone in oocyte acquisition of developmental competence. *Reprod. Domest. Anim.* **47**, 142–147 (2012). Suppl 4.
- Hynes, N. E. & Lane, H. A. ERBB receptors and cancer: the complexity of targeted inhibitors. *Nat. Rev. Cancer* **5**, 341–354 (2005).
- Bagnoli, M. et al. Choline metabolism alteration: a focus on ovarian cancer. *Front. Oncol.* **6**, 153 (2016).
- Rosenzweig, A., Blenis, J. & Gomes, A. P. Beyond the Warburg effect: how do cancer cells regulate one-carbon metabolism? *Front Cell Dev. Biol.* **6**, 90 (2018).
- Hopper-Borge, E. A. et al. Mechanisms of tumor resistance to EGFR-targeted therapies. *Expert Opin. Ther. Targets* **13**, 339–362 (2009).
- Matei, D., Chang, D. D. & Jeng, M. H. Imatinib mesylate (Gleevec) inhibits ovarian cancer cell growth through a mechanism dependent on platelet-derived growth factor receptor α and Akt inactivation. *Clin. Cancer Res.* **10**, 681–690 (2004).
- Apte, S. M., Bucana, C. D., Killion, J. J., Gershenson, D. M. & Fidler, I. J. Expression of platelet-derived growth factor and activated receptor in clinical specimens of epithelial ovarian cancer and ovarian carcinoma cell lines. *Gynecol. Oncol.* **93**, 78–86 (2004).
- Schmitt, J. & Matei, D. Platelet-derived growth factor pathway inhibitors in ovarian cancer. *Clin. Ovarian Cancer Other Gynecol. Malig.* **1**, 120–126 (2008).
- Boctor, A. et al. Imatinib in refractory crohn disease: a series of 6 cases. *Crohn's Colitis* **360**, 1 (2019).
- Can, G. et al. The Syk inhibitor fostamatinib decreases the severity of colonic mucosal damage in a rodent model of colitis. *J. Crohns. Colitis* **9**, 907–917 (2015).
- Tigno-Aranjuez, J. T., Asara, J. M. & Abbott, D. W. Inhibition of RIP2's tyrosine kinase activity limits NOD2-driven cytokine responses. *Genes Dev.* **24**, 2666–2677 (2010).
- Overstreet, A. M. et al. The JAK inhibitor ruxolitinib reduces inflammation in an ILC3-independent model of innate immune colitis. *Mucosal Immunol.* **11**, 1454–1465 (2018).
- Stelzer, G. et al. The GeneCards suite: from gene data mining to disease genome sequence analyses. *Curr. Protoc. Bioinforma.* **54**, 1.30.1–1.30.33 (2016).
- Fay, W. P., Parker, A. C., Condrey, L. R. & Shapiro, A. D. Human plasminogen activator inhibitor-1 (PAI-1) deficiency: characterization of a large kindred with a null mutation in the PAI-1 gene. *Blood* **90**, 204–208 (1997).
- Jankun, J. et al. Highly stable plasminogen activator inhibitor type one (VLHL PAI-1) protects fibrin clots from tissue plasminogen activator-mediated fibrinolysis. *Int. J. Mol. Med.* **20**, 683–687 (2007).
- Hodges, A. et al. Regional and cellular gene expression changes in human Huntington's disease brain. *Hum. Mol. Genet.* **15**, 965–977 (2006).
- Jones, L. et al. Assessment of the relationship between pre-chip and post-chip quality measures for Affymetrix GeneChip expression data. *BMC Bioinforma.* **7**, 211 (2006).
- Rabbani, G. H., Greenough, W. B. 3rd, Holmgren, J. & Lönnroth, I. Chlorpromazine reduces fluid-loss in cholera. *Lancet* **1**, 410–412 (1979).
- Beister, A. et al. The N-methyl-D-aspartate antagonist memantine retards progression of Huntington's disease. *J. Neural Transm. Suppl.* **68**, 117–122 (2004).
- Shen, Y.-C. Lamotrigine in motor and mood symptoms of Huntington's disease. *World J. Biol. Psychiatry* **9**, 147–149 (2008).
- Vattakatuchery, J. J. & Kurien, R. Acetylcholinesterase inhibitors in cognitive impairment in Huntington's disease: A brief review. *World J. Psychiatry* **3**, 62–64 (2013).
- Murray, T. F., Mpitsos, G. J., Siebenaller, J. F. & Barker, D. L. Stereoselective L-[3H]quinuclidinyl benzilate-binding sites in nervous tissue of *Aplysia californica*: evidence for muscarinic receptors. *J. Neurosci.* **5**, 3184–3188 (1985).
- Murkin, L., Hussain, K. & Schilder, A. G. M. Betahistine for symptoms of vertigo. *Cochrane Database Syst. Rev.* CD010696 (2016).
- De Marchi, N., Daniele, F. & Ragone, M. A. Fluoxetine in the treatment of Huntington's disease. *Psychopharmacology* **153**, 264–266 (2001).
- Li, S. & Yang, J. Pitolisant for treating patients with narcolepsy. *Expert Rev. Clin. Pharmacol.* **13**, 79–84 (2020).
- Cummings, J., Lee, G., Ritter, A., Sabbagh, M. & Zhong, K. Alzheimer's disease drug development pipeline: 2019. *Alzheimers Dement.* **5**, 272–293 (2019).

50. Wharton, W. et al. Rationale and design of the mechanistic potential of antihypertensives in preclinical Alzheimer's (HEART) trial. *J. Alzheimers Dis.* **61**, 815–824 (2018).
51. Davies, N. M., Kehoe, P. G., Ben-Shlomo, Y. & Martin, R. M. Associations of anti-hypertensive treatments with Alzheimer's disease, vascular dementia, and other dementias. *J. Alzheimers Dis.* **26**, 699–708 (2011).
52. Li, N.-C. et al. Use of angiotensin receptor blockers and risk of dementia in a predominantly male population: prospective cohort analysis. *BMJ* **340**, b5465 (2010).
53. Guimond, M.-O. & Gallo-Payet, N. The Angiotensin II type 2 receptor in brain functions: an update. *Int. J. Hypertens.* **2012**, 351758 (2012).
54. Butterfield, D. A. & Halliwell, B. Oxidative stress, dysfunctional glucose metabolism and Alzheimer disease. *Nat. Rev. Neurosci.* **20**, 148–160 (2019).
55. Kuehn, B. M. In Alzheimer research, glucose metabolism moves to center stage. *JAMA* **323**, 297–299 (2020).
56. Yarchoan, M. & Arnold, S. E. Repurposing diabetes drugs for brain insulin resistance in Alzheimer disease. *Diabetes* **63**, 2253–2261 (2014).
57. Claxton, A. et al. Long-acting intranasal insulin detemir improves cognition for adults with mild cognitive impairment or early-stage Alzheimer's disease dementia. *J. Alzheimers Dis.* **44**, 897–906 (2015).
58. Rosenbloom, M. H. et al. A phase II, single center, randomized, double-blind, placebo-controlled study of the safety and therapeutic effectiveness of intranasal glulisine in amnesic mild cognitive impairment and probable mild Alzheimer's disease: Human/Human trials: Other. *Alzheimers. Dement.* **16**, e036840 (2020).
59. Di Paolo, G. & Kim, T.-W. Linking lipids to Alzheimer's disease: cholesterol and beyond. *Nat. Rev. Neurosci.* **12**, 284–296 (2011).
60. Masters, C. L. et al. Alzheimer's disease. *Nat. Rev. Dis. Prim.* **1**, 15056 (2015).
61. Seddighi, S., Varma, V. & Thambisetty, M. α 2-macroglobulin in Alzheimer's disease: new roles for an old chaperone. *Biomark. Med.* **12**, 311–314 (2018).
62. Giri, M., Shah, A., Upreti, B. & Rai, J. C. Unraveling the genes implicated in Alzheimer's disease. *Biomed. Rep.* **7**, 105–114 (2017).
63. Chandra, S. & Pahan, K. Gemfibrozil, a lipid-lowering drug, lowers amyloid plaque pathology and enhances memory in a mouse model of Alzheimer's disease via peroxisome proliferator-activated receptor α . *J. Alzheimers Dis. Rep.* **3**, 149–168 (2019).
64. Stibius, K. B. & Snieppen, K. Modeling the two-hybrid detector: experimental bias on protein interaction networks. *Biophys. J.* **93**, 2562–2566 (2007).
65. Schaefer, M. H., Serrano, L. & Andrade-Navarro, M. A. Correcting for the study bias associated with protein-protein interaction measurements reveals differences between protein degree distributions from different cancer types. *Front. Genet.* **6**, 260 (2015).
66. Brown, A. S. & Patel, C. J. A standard database for drug repositioning. *Sci. Data* **4**, 170029 (2017).
67. Kou, L., Markowsky, G. & Berman, L. A fast algorithm for Steiner trees. *Acta Inf.* **15**, 141–145 (1981).
68. Järvelin, K. & Kekäläinen, J. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst. Secur.* **20**, 422–446 (2002).

Acknowledgements

S.S., J.S., E.A., K.F., S.C., H.H.H.W.S., J.Ba., A.W., and T.K. are grateful for financial support from REPO-TRIAL. REPO-TRIAL has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777111. This publication reflects only the authors' view and the European Commission is not responsible for any use that may be made of the information it contains. J.Ba. and T.K. are grateful for financial support from BMBF grant Sys_CARE (no. 01ZX1908A) of the Federal German Ministry of Research and Education. J.Ba. was partially funded by

his VILLUM Young Investigator Grant no. 13154. Contribution by J.Be. is funded by the German Federal Ministry of Education and Research (BMBF) within the framework of the e:Med research and funding concept (grant 01ZX1910D). M.S.-A. is grateful for a Ph.D. fellowship funding from CONACYT (CVU659273) and the German Academic Exchange Service, DAAD (ref. 91693321). Contribution by O.L. is funded by the Bavarian State Ministry of Science and the Arts as part of the Bavarian Research Institute for Digital Transformation. A.I.C. is currently financially supported by the DFG Walter Benjamin Program (ref. DFG CA 2642/1-1). Figures 1 and 2 are created with BioRender.com.

Author contributions

S.S., J.S., D.B.B., J.Ba., A.W., and T.K. conceived the idea and designed the platform. S.S., J.S., J.Be., E.A., G.G., K.F., S.C., T.K. performed the acquisition, harmonization and integration of databases. S.S. and D.B.B. developed and adapted the network-based algorithms for drug repurposing. S.S., E.A., G.G., M.S.-A., O.L., C.N., and A.I.C. discovered and approved the use cases. J.S. implemented the API. S.S. and J.Be. implemented the Cytoscape app. All authors provided critical feedback and discussion, assisted in the interpretation of data and use cases, writing the manuscript, and the improvement of the platform.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-021-27138-2>.

Correspondence and requests for materials should be addressed to Sepideh Sadegh.

Peer review information *Nature Communications* thanks Lincoln Stein and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at <http://www.nature.com/reprints>

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



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

© The Author(s) 2021

A.3. Lessons from the COVID-19 pandemic for advancing computational drug repurposing strategies



Lessons from the COVID-19 pandemic for advancing computational drug repurposing strategies

Gihanna Galindez^{1,5}, Julian Matschinske^{1,5}, Tim Daniel Rose^{2,5}, Sepideh Sadegh^{1,5},
Marisol Salgado-Albarrán^{1,3,5}, Julian Späth^{1,5}, Jan Baumbach^{1,4,6} and Josch Konstantin Pauling^{1,6}✉

Responding quickly to unknown pathogens is crucial to stop uncontrolled spread of diseases that lead to epidemics, such as the novel coronavirus, and to keep protective measures at a level that causes as little social and economic harm as possible. This can be achieved through computational approaches that significantly speed up drug discovery. A powerful approach is to restrict the search to existing drugs through drug repurposing, which can vastly accelerate the usually long approval process. In this Review, we examine a representative set of currently used computational approaches to identify repurposable drugs for COVID-19, as well as their underlying data resources. Furthermore, we compare drug candidates predicted by computational methods to drugs being assessed by clinical trials. Finally, we discuss lessons learned from the reviewed research efforts, including how to successfully connect computational approaches with experimental studies, and propose a unified drug repurposing strategy for better preparedness in the case of future outbreaks.

The novel SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2) pathogen has infected around 60 million people and caused more than a million deaths worldwide (<https://covid19.who.int/>; as of November 2020). As a result, there is a need to find treatments that can be applied immediately to reduce mortality or morbidity.

Repurposing existing drugs is a rapid and effective way to provide such treatments by identifying new uses for drugs that have well-established pharmacological and safety profiles¹. Many drugs used to treat different diseases have already been successfully repurposed and approved for new indications². While repurposing can be conducted at any point in drug development, its greatest potential can be applied to drugs that are already approved³. In the case of the COVID-19 pandemic, it is a fast and cost-efficient approach to identify novel treatments⁴.

Recent studies have increasingly employed computational methods to systematically predict new drug targets or drug repurposing candidates. In contrast to experimental high-throughput screening, *in silico* approaches are faster, lower-cost, and can serve as an initial filtering step for evaluating thousands of compounds. Thus, they are useful for prioritizing drugs that warrant further evaluation and experimental validation. This requires the application of suitable algorithmic approaches to identify mechanisms relevant or specific to the disease⁴.

This Review discusses current *in silico* drug repurposing efforts for COVID-19, followed by a discussion of the lessons learned from different perspectives (from data resources to the quality of predictions) and a proposed unified strategy to improve the response in

potential future outbreaks. The covered studies employed standard drug repurposing workflows and data-driven algorithms.

As new studies are published almost every day, it is not possible to provide a broad and comprehensive overview of all repurposing studies. Hence, this Review focuses on the computational methods for drug repurposing, their application, availability and feasibility in a selection of studies (peer-reviewed and preprint) that were selected to cover a wide variety of different methods. It is worth noting that most of these studies are not considered successful clinically. Nevertheless, it is important to properly evaluate and improve the predictive power of *in silico* approaches that are capable of utilizing information from existing drugs as well as host and virus biology, even with limited availability of data on the novel emerging pathogen. This promotes a rapid and practical response to infection and therefore improves success in future pandemics, particularly in tackling the rise in infection cases at the early stages of the pandemic or ahead of vaccine development.

Data resources

Besides experimental datasets, the rapid availability of resources that integrate different data types is crucial in a pandemic. Sharing data accelerates research, as computational methods depend on high-quality datasets, and experimental labs do not need to collect the information on their own. The large number of resources used in COVID-19 drug repurposing studies have shown that data can be quickly generated and gathered through strong community efforts. This section presents a selection of data resources used in the reviewed studies to describe the resource types that accelerated

¹Chair of Experimental Bioinformatics, TUM School of Life Sciences, Technical University of Munich, Munich, Germany. ²LipiTUM, Chair of Experimental Bioinformatics, TUM School of Life Sciences, Technical University of Munich, Munich, Germany. ³Natural Sciences Department, Universidad Autónoma Metropolitana-Cuajimalpa (UAM-C), Mexico City, Mexico. ⁴Computational Biomedicine Lab, Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark. ⁵These authors contributed equally: Gihanna Galindez, Julian Matschinske, Tim Daniel Rose, Sepideh Sadegh, Marisol Salgado-Albarrán, Julian Späth. ⁶These authors jointly supervised this work: Jan Baumbach, Josch Konstantin Pauling.

✉e-mail: josch.pauling@wzw.tum.de

computational drug repurposing approaches: most of them are general data resources that were already established before the pandemic but that have been extended with COVID-19 or SARS-CoV-2-specific data. The resources used in the reviewed studies are listed in Supplementary Table 1. A list of COVID-19 specific data resources that were not used in the reviewed studies but may become relevant in the future is given in Supplementary Table 2.

Molecular data resources. All molecular data used in the reviewed publications were extracted from already established, general data resources that were quickly extended with SARS-CoV-2-specific data. Resources such as GenBank⁵, the GISAID initiative⁶, or UniProt⁷ provide genomic/proteomic sequence information about hosts and SARS-CoV-2. Structural resources collecting information about proteins, such as the Protein Data Bank (PDB)⁸, were extended by various SARS-CoV-2-specific proteins. Finally, transcriptome resources that collect gene expression data were used in several COVID-19 drug repurposing approaches. For instance, the Genotype-Tissue Expression (GTEx)⁹ program offers insights into tissue-specific gene expression. Expression in lung tissues is of high interest in COVID-19 drug repurposing research and was often integrated in computational models or studies. Other resources, such as the LINCS L1000 database¹⁰, profile gene expression changes under certain drug treatment conditions and were used to identify drugs with reverse expression profiles to the samples infected with SARS-CoV-2.

Network and interaction resources. Protein–protein interaction (PPI) networks enable visualization and analyses of the interactions between either host or virus proteins and other host proteins. Furthermore, PPI networks allow for particular adaptation and search strategies (for example, edge filtering) and can be connected to drug resources. Gordon et al.¹¹ identified 332 high-confidence virus–host interactions between SARS-CoV-2 and human proteins. It was the only newly created and exclusively SARS-CoV-2-related resource used in the reviewed publications of this work. VirHostNet^{12,13}, a virus–host PPI resource that already existed before the 2019/2020 SARS outbreak, was expanded with 167 new SARS-CoV-2 interactions. In contrast to virus–host PPIs, host PPIs are not virus specific. All resources that were used in the reviewed studies were already available before the pandemic but have since been widely used in COVID-19 drug repurposing approaches^{14,15}. Besides molecular networks, knowledge graphs, such as the Global Network of Biomedical Relationships (GNBR)¹⁶, have demonstrated their utility for drug repurposing. These networks comprise various types of biological relationships assembled from literature and were integrated into COVID-19 drug repurposing approaches¹⁷.

Drug and trial resources. Drug databases that already existed before the pandemic and that are continuously extended with newly developed drugs were used to connect the results of different approaches to potential drugs. A widely used drug database is DrugBank¹⁸, with more than 13,000 drug entries of approved and in-trial drugs, including drug targets. On the other hand, ChEMBL¹⁹ and ZINC15²⁰ contain millions of compounds that exhibit drug-like properties.

Drug repurposing approaches also benefited from trial databases as they can be used to validate whether the predicted drugs are already in trial or have not yet been evaluated. Examples of such resources are the EU Clinical Trials Register (<https://www.clinicaltrialsregister.eu/>) and ClinicalTrials.gov (<https://clinicaltrials.gov/>). The latter contains more than 350,000 research studies from 219 countries.

Drug repurposing studies

Various clinical, experimental and computational drug repurposing efforts have been rapidly mobilized prioritizing compounds to

identify promising drug candidates for the SARS-CoV-2 pandemic. In this section, we examine a selection of studies representing the different computational approaches to identify potential new targets and repurposable drugs for COVID-19.

Virus-targeting approaches. Virus-targeting approaches mostly rely on structure-based drug screening methods, which take the three-dimensional structures of target proteins to predict affinities or interaction energies of known chemical compounds to the proteins (Fig. 1). These methods were mainly used to identify candidate drugs that target viral proteins, so we refer to them as virus-targeting approaches, although they can also be applied to host proteins. Two main methodological workflows were applied, namely, structure-based²¹ and deep-learning (DL)-based drug screening. Here, we describe these methods and compare 23 COVID-19 drug repurposing studies^{22–44}.

Structure-based drug screening. The first step for structure-based screening is the selection of the drug library and the target protein. For COVID-19, the intuitive candidate for targeting virus proteins were antivirals. Thus, many studies limited their search to these. The number of screened drugs ranged from 3 (ref.³⁷) to 123 antiviral drugs³³. Broader studies, such as that by Chen et al.²⁶, combined compounds from the KEGG (Kyoto Encyclopedia of Genes and Genomes) and DrugBank databases to screen 7,173 drugs.

The other crucial step is the selection of the target protein and its corresponding three-dimensional structure (experimental or predicted). Wu et al.⁴⁰ performed screening on 19 encoded proteins of the virus. By comparison, most other studies focused on the 3CLpro, envelope (E), spike, RNA polymerase and methyltransferase proteins.

Virtual screening of the drug libraries utilized established software, such as Autodock⁴⁵ and Glide⁴⁶. Candidate drugs were selected using respective scoring methods, followed by validations with molecular dynamics simulations^{30,37}.

Most drugs were predicted for 3CLpro (Supplementary Table 3), which was also the focus of most studies (17 studies), followed by RdRp and PLPro. For 3CLpro, the predictions ranged from 2 (ref.²⁹) to 27 (ref.⁴⁰) drugs per study. The 5 most frequently predicted drugs were ritonavir (8 studies), lopinavir (6 studies), nelfinavir, remdesivir and saquinavir (5 studies each). However, 99 of the candidate drugs were only predicted in 1 study, showing a high variability in the resulting candidate sets. Interestingly, the studies that screened full databases also predicted antiviral drugs as top scorers (Supplementary Table 4). Of the 23 studies, 10 have not yet been peer-reviewed, which we discuss in the section on ‘A unified drug repurposing strategy’.

DL-based repurposing strategies. DL models can predict binding affinities or docking scores and have shown advantages over conventional docking protocols. While standard docking protocols are limited to millions⁴⁷, DL approaches can analyze billions of chemical compounds. This allows them to be applied to whole databases, which increases the diversity of the tested compounds and the likelihood of finding unconventional compounds⁴⁷. Furthermore, they are capable of processing more (physico-)chemical features⁴⁸ and can find features related to a non-favorable docking⁴⁷. However, most of these methods require datasets for training, which often come from real docking simulations; thus, the performance of many DL-based approaches still rely on the accuracy of the docking software used for training.

Ton et al.⁴² developed DeepDocking⁴⁷, which utilizes quantitative structure–activity relationship models trained to predict docking scores of compounds targeting the SARS-CoV-2 3CLpro protein. It requires fewer docking pipelines, since it performs docking only

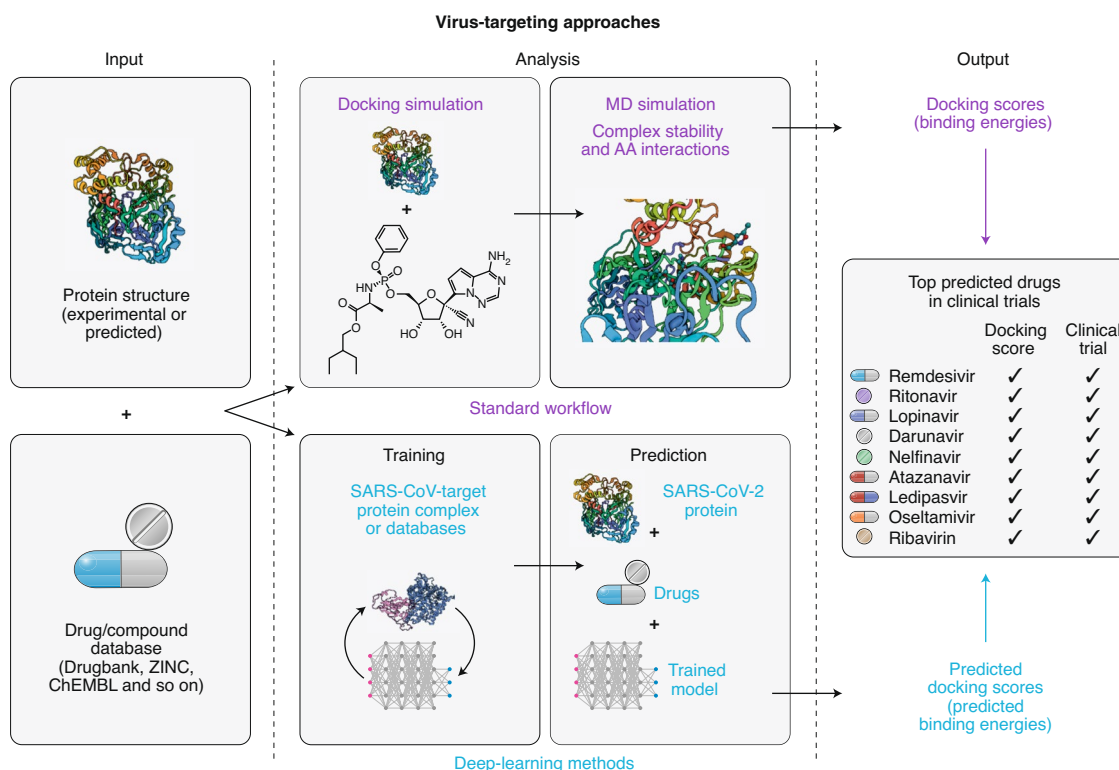


Fig. 1 | Workflows of virus-targeting computational drug repurposing approaches. The input data consist of protein structure information (experimental or predicted) and chemical structure of drugs from public databases. Two analysis workflows can be applied: standard analysis consisting of docking followed by molecular dynamics (MD) simulations and DL-based analysis. Finally, the output data of both approaches generally consist of a ranking of drugs based on their (predicted) docking scores. The drugs can be further evaluated by whether or not they are in clinical trials.

on subsets of compounds and can produce a reduced list of compounds, which is also enriched in potential top hits.

Nguyen et al.⁴⁹ developed the method MathDL, which utilizes low-dimensional mathematical representations of the drug–target protein complex structures, which are then fed to DL algorithms to predict binding energies of drug–protein complexes. For SARS-CoV-2, the authors used experimental binding affinity data from SARS-CoV ligand–3CLpro complexes from PDBbind and SARS-CoV protease inhibitors as training data to predict binding energies on DrugBank compounds for SARS-CoV-2 3CLpro (ref.⁵⁰) and does not depend on docking software.

Beck et al.⁴⁴ developed a DL-based drug–target interaction prediction model, named Molecule Transformer-Drug Target Interaction. It utilizes simplified molecular-input line-entry system (SMILES)⁵¹ representations for drugs and protein sequences as input for training and predicts affinities. For SARS-CoV-2, the model was trained on commercially available antiviral drugs and viral target proteins. Antiviral drugs already used against SARS-CoV-2 were found among the candidate drugs identified.

Host-targeting approaches. Host-targeting approaches involve identifying potential drugs that interfere with host mechanisms that contribute to viral pathogenesis, which also makes them less prone to drug resistance^{52,53}. In addition, SARS-CoV-2 infections can trigger a hyper-reactive immune response characterized by the excessive release of pro-inflammatory cytokines and chemokines⁵⁴. Thus, drugs that modulate the host immune response can benefit critically ill patients with COVID-19 by targeting specific dysregulated pathways^{54–56}.

Signature-based approaches. Signature-based approaches primarily utilize transcriptome datasets from samples infected with SARS-CoV-2 or closely related human coronaviruses to identify

candidate drugs through connectivity mapping (Fig. 2), a well-established approach that relies on finding drug-induced expression signatures exhibiting reverse profiles to a disease signature^{57,58}. Several studies adopted this as a primary method for identifying new therapeutics for COVID-19. Loganathan et al.⁵⁹ performed differential expression analysis of virus-infected cells and extracted consistently dysregulated genes in infected conditions. They were used to query the Connectivity Map database⁵⁸ for drug perturbation profiles exhibiting anti-correlated expression signatures. A modified approach was implemented by Jia et al.⁶⁰, wherein expression data from infected and healthy individuals were used as input to a pathway-guided drug repurposing framework. They identified disease co-expression clusters and performed enrichment analyses prior to reverse signature matching⁶⁰.

Network-based approaches. The general network-based approach applied in drug repurposing studies on COVID-19 integrates multiple data sources, including virus–host interactions, PPIs, co-expression networks, functional associations or drug–target interactions (Fig. 2). Network-based algorithms or topology measures are applied to the assembled networks to identify relevant host protein targets or regions of the host interactome that can be targeted.

Multiple studies implement random-walk-based algorithms as the primary method to identify new putative drug targets. Law et al.⁶¹ implemented several algorithms on a virus–host interactome to identify additional SARS-CoV-2 interactors. The coronavirus spike protein primarily has been established to mediate viral entry into host cells⁶². Similarly, but focusing on a specific context, Messina et al.⁶³ explored the pathogenic mechanisms triggered by the spike protein using data from three closely related coronaviruses. They

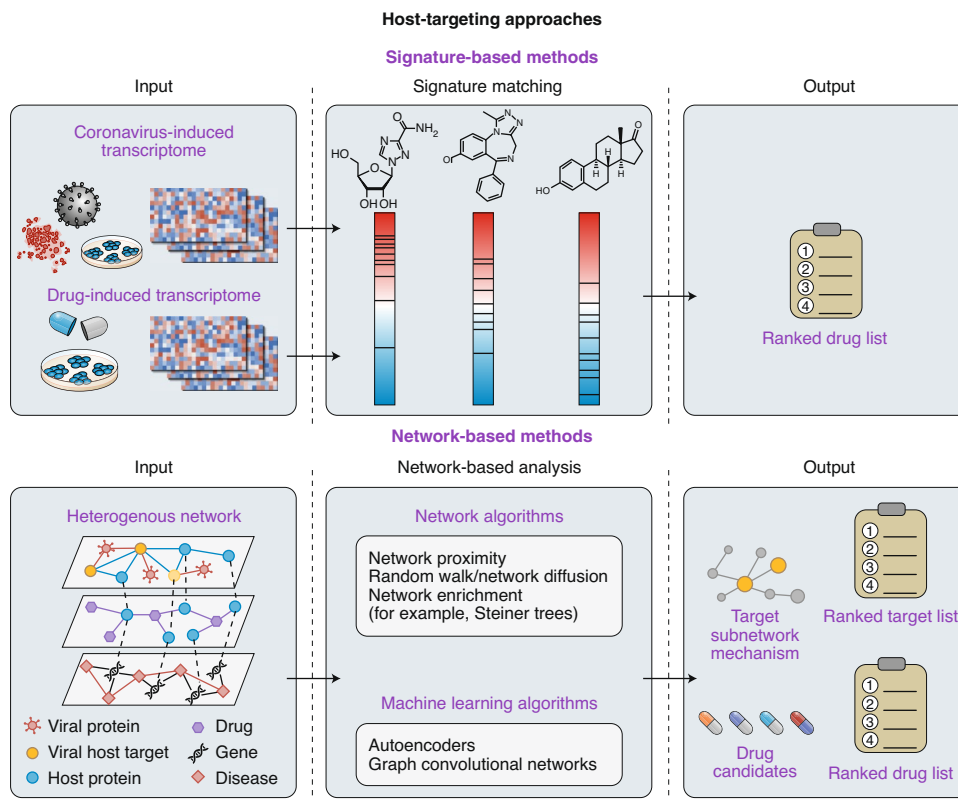


Fig. 2 | Workflows of host-targeting computational drug repurposing approaches. Signature-based methods involve finding drug-induced expression profiles that exhibit reverse patterns to the coronavirus disease signature. Network-based approaches typically assemble heterogeneous networks from diverse data types, including gene–disease associations or drug–target associations. Algorithms such as network proximity, random walk/diffusion-based methods, or network enrichment are then employed. Some studies combined them with machine-learning-based methods, particularly autoencoders and graph convolutional networks. The outputs can be ranked lists of host targets or drug candidates.

implemented a random walk algorithm on assembled molecular networks using the spike protein as seed to identify relevant targets for COVID-19⁶³. In addition, CoVex⁶⁴ implemented TrustRank⁶⁵, a variant of the PageRank⁶⁶ algorithm, to propagate scores from user-defined seeds to the other host proteins and rank host drug targets.

Network proximity relies on the principle that a drug can be effective if it targets proteins within the neighborhood of disease-associated proteins in the interactome⁶⁷. Zhou et al.⁶⁸ utilized this concept to compute the network proximity measure between drug targets and coronavirus-associated proteins in the human interactome. They also used the ‘complementary exposure’ pattern, which is based on the shortest distance between targets of two drugs predicted by network proximity, to identify potential drug combinations to treat COVID-19 patients⁶⁸.

Several studies combined multiple network-based strategies to predict drug candidates. Gysi et al.⁶⁹ characterized and extracted a COVID-19 disease module using experimentally determined SARS-CoV-2 interactors. They performed network-based analyses accounting for tissue specificity and potential disease comorbidities. They employed a multi-modal approach to the virus–host interactome integrating network proximity, diffusion state distance and graph convolutional networks (GCNs) to identify drugs that can perturb the activity of host proteins associated with the COVID-19 disease module. The final drug list was obtained by rank aggregation from the different pipelines⁶⁹.

CoVex⁶⁴ is a web platform for exploring SARS-CoV and SARS-CoV-2 virus–host–drug interactomes⁶⁴. Users can predict drug targets and drug candidates using several graph analysis methods that allow custom seed proteins as input. For instance,

KeyPathwayMiner⁷⁰ is a network enrichment tool that identifies condition-specific subnetworks by extracting a maximally connected subnetwork from the host interactome starting from the seeds. CoVex also implements a weighted multi-Steiner tree method that aggregates several non-unique approximations of Steiner trees, which are subnetworks of minimum cost connecting the set of seeds, into a single subnetwork.

Other studies additionally utilize machine learning to predict drug candidates against SARS-CoV-2. Belyaeva et al.⁷¹ implemented a hybrid approach between signature matching and network-based methods. Using autoencoders, they learned feature embeddings for drugs using drug-induced expression profiles to identify drugs exhibiting reverse profiles to the SARS-CoV-2 infection signature. Steiner tree and causal network discovery algorithms were then used to extract the mechanisms mediated by both SARS-CoV-2 and aging⁷¹. Ge et al.⁷² constructed a virus-related knowledge graph and employed a GCN algorithm. The list of drug candidates was further filtered for existing evidence of antiviral activities through text mining⁷². Similarly, Zeng et al.¹⁷ assembled a large-scale knowledge graph derived from PubMed articles. A GCN model was then applied to learn low-dimensional embeddings of the nodes and edges¹⁷.

Lessons learned

In the following, we examine the quality and potential of the reviewed data resources and computational methods in order to improve the response in future pandemics.

Data resources. The availability of molecular datasets is a precondition to develop drug repurposing methods quickly. Besides that, network-based resources were a large driver in

drug repurposing. However, a large portion of the publications are based on only a few primary resources, which always induces the risk of bias or measurement errors. In addition, the only type of molecular interaction network used was PPI. Still, high confidence PPIs are needed since, for instance, none of the approaches included structure data. In the future, other network types, such as gene regulatory networks, should be considered. Other data resources, such as off-label data for drugs, should also be integrated in drug repurposing studies.

Finally, existing drug and trial resources were widely used for developing the drug repurposing pipelines. However, we observed no standardization in trial resources, making it hard to analyze trials for certain drugs due to different names, different spellings, or typing errors. Standardization is usually implemented for drug resources (for example, DrugBank), but some drugs undergoing trials could not be found in the databases. Keeping the resources up to date and interconnected should be a focus and will enhance accessibility.

Computational predictions. Assessing the quality of predictions is challenging, since many studies are not peer-reviewed, do not perform experimental evaluation, or rely on clinical trial databases. We examined the quality of predictions by determining the overlap between the final candidate drug lists from the individual studies and the drugs undergoing clinical trials from ClinicalTrials.gov (<https://clinicaltrials.gov/>) and Biorender (<https://biorender.com/covid-vaccine-tracker>) databases. In addition, we provide supplementary in vitro screening data, such as IC₅₀ values for viral targets and inhibition indices from cell culture studies for SARS-CoV-2 (Supplementary Data 1). Our effort to compile these data shows that a substantial number of predictions have not been experimentally tested.

Evaluating virus-targeting approaches. We identified 53 drugs predicted with docking simulations that are undergoing current trials (Supplementary Table 5). Wu et al.⁴⁰ identified most of the drugs (36 drugs); however, these drugs were predicted for multiple viral proteins (for example, chlorhexidine for 11 and methotrexate for 6 different viral proteins). This indicates that their approach did not yield specific and feasible candidates. After excluding this study, the remaining drugs were only predicted for one specific protein each, except for chloroquine (3CLpro and PLpro) and remdesivir (3CLpro and RdRp). The top five drugs in clinical trials, which were predicted by docking simulations using the 3CLpro main protease, were predicted by 14.3% (darunavir), 19.0% (remdesivir), and 23.8% (lopinavir, nelfinavir, ritonavir) of the total number of included docking studies (Supplementary Table 6), showing that for each drug, the majority of studies were not able to predict them. Similar drugs were identified by the DL approach of Beck et al.⁴⁴, who identified ritonavir, lopinavir and remdesivir, which are being tested in multiple clinical trials. However, these antiviral drugs have not yet shown well-defined results in patients. For ritonavir/lopinavir, only four trials are completed^{73–76} and preliminary results suggest no difference in the outcome after treatment^{77–79}. Further investigation is required⁸⁰. For remdesivir, some trials have been completed and the preliminary results in patients^{81–83} and human cell lines⁸⁴ showed that it could be effective in treating SARS-CoV-2 infection.

Antiviral drugs are always the top hits among a large selection of drugs from databases, indicating high accuracy of the methods. These drugs are good candidates for experimental screening or clinical trials, independently of how reliable the computational predictions are. More interesting candidates are the additional drugs identified by these approaches; however, little experimental validation is available for these drugs and the majority of them do not enter clinical trials. A similar situation is observed in the emerging

field of DL approaches, where most studies focused on demonstrating the accuracy of their predictions and developing benchmarking datasets^{85,86}. DL and docking simulation-based approaches are promising tools to identify repurposable drugs given their capacity to deliver results in a short time. While a standard workflow is already established for docking simulations, DL-based approaches might robustly deliver testable candidate drugs. However, docking studies in particular were rarely peer reviewed, found very different candidate sets and partially used different scores for evaluation and ranking. This makes it necessary to validate these results by systematic comparisons of experiments.

Evaluating host-targeting approaches. Host-targeting approaches typically involve integration and analysis of multiple omics types and employ data-driven network-based methods; thus, a major limitation is the lack of gold-standard datasets and the scarcity of data from the MERS-CoV (Middle East respiratory syndrome coronavirus) and SARS-CoV outbreaks. Prior to the availability of sufficient SARS-CoV-2-specific data, earlier studies utilized preliminary data or augmented the analyses using data from closely related viruses. While the quality of the predictions is highly data-dependent, continued generation of SARS-CoV-2-specific omics data and pending results on clinical studies are expected to improve the predictions. Clinical expert knowledge remains crucial for filtering the drug predictions based on criteria such as toxicity and pharmacological properties. However, the efficacy of these candidate drugs in trial remains to be established and firm conclusions cannot be made because of the limited data availability.

The degree of overlap with drugs in clinical trials was generally low (Supplementary Tables 7 and 8), but more than half of the drugs (26 out of 41) predicted by an ensemble method primarily based on knowledge graphs¹⁷ are also undergoing clinical trials. While it should be noted that the drugs registered for clinical trials were also used as their validation set at the time of writing, more of their predicted drugs were registered for clinical trials later on. We also noted several drugs that were predicted by both signature-based and network-based approaches and thus warranted further examination (Supplementary Table 9). Ribavirin was predicted by four out of six studies^{17,60,69,71}, thereby providing a mechanistic basis for its predicted efficacy. Methotrexate, which is indicated for rheumatoid arthritis, was also predicted by three studies^{17,68,69}.

It is worth noting that several predicted compounds are currently used to treat critically ill COVID-19 patients. An example is dexamethasone (predicted by one signature-based⁶⁰ and two network-based studies^{17,69}), which was supported by the RECOVERY trial⁸⁷. Hydrocortisone (predicted by three studies^{17,68,69}) has also demonstrated efficacy for critically ill patients⁸⁸. Dexamethasone and hydrocortisone are corticosteroids that act by modulating an overactive immune response, which is typically observed in severely ill COVID-19 patients.

Notably, drugs reaching advanced phases in clinical trials were not selected based on in silico predictions, but were repurposed based on clinical experience with the previous SARS or MERS outbreaks⁸⁹ and selected based on known effects in alleviating disease symptoms. Furthermore, the predictions were not followed-up by experimental validation in the majority of the studies reviewed. This translational gap between computational efforts for drug repurposing and clinical application is a major and widely recognized bottleneck in drug repurposing and medicine in general. Results from systematic validation efforts will also be important for identifying the algorithms and datasets that are specifically suitable for drug repurposing in the COVID-19 context. Given the urgency of identifying effective therapies in a pandemic, close collaboration between clinicians, experimental biologists and computational biologists is expected to address this gap.

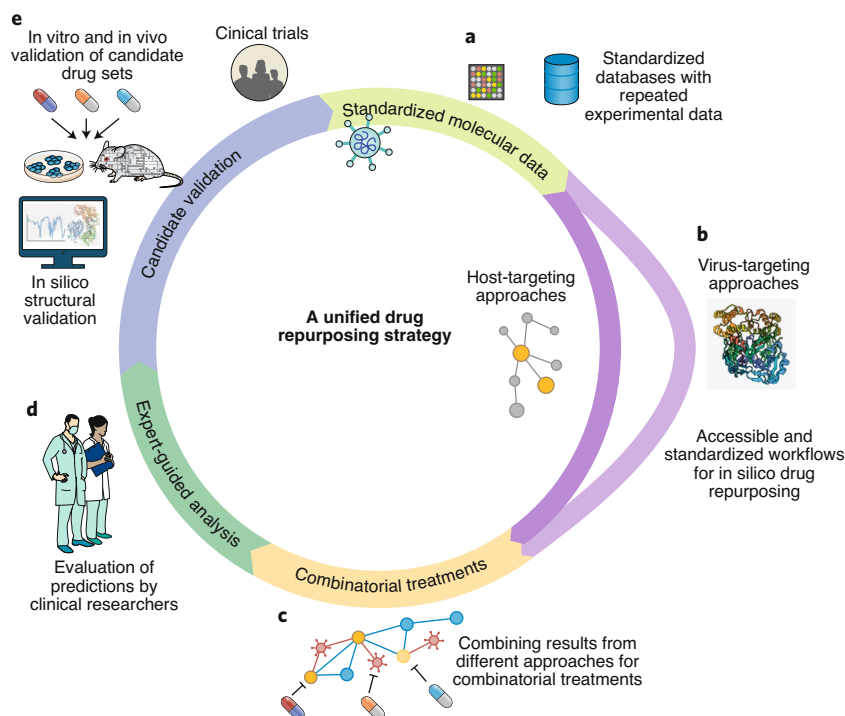


Fig. 3 | Proposed elements of a unified drug repurposing strategy. **a**, Availability of standardized data. **b**, Accessible workflows for computational predictions. **c**, Combination of predictions from different methods. **d**, Feedback from clinical experts of drug candidate sets and screening parameters. **e**, Validation of predicted drugs with different approaches.

A unified drug repurposing strategy. Although overlaps between computationally predicted drug repurposing and clinical trials exist, there are no indications that clinical trials were conducted based on computational predictions, despite their promising potential. For future pandemics, computational tools should be able to deliver promising sets of candidates, which could then be validated in trials or screenings. Therefore, a unified strategy is necessary. In the following, we identify important issues and discuss potential solutions to make computational drug repurposing more effective.

Availability of standardized data. Newly developed methods often rely on the same data types (Fig. 3a). The fast generation of different kinds of data in future disease outbreaks is a key initial step. Notable examples are the interaction data from Gordon et al.¹¹ and the publication of the 3CLpro⁹⁰ structure, which were both used by many subsequent studies. However, experimental replication of datasets obtained from different laboratories and the integration of different data types are crucial to increase robustness and require improvement.

Tool accessibility. Despite the large variety of computational tools and software, it has so far been of limited practical use to clinical researchers during the COVID-19 pandemic (Fig. 3b). For virus-targeting therapies, docking pipelines remain stable and a large amount of software has been developed; however, their corresponding outputs showed wide variability depending on the algorithm used, lowering comparability (standardization problem). For host-targeting therapies, the in silico pipelines are more methodologically diverse and several strategies were developed to target specific biological contexts. However, the general availability of computational tools and software in the context of the COVID-19 pandemic has been highly limited. Tool accessibility allows researchers to run custom analyses using the developed algorithms (for example, on newly available data). This will help non-computational scientists to use these tools and continue with validation routines, avoiding many

preprint manuscripts that are never validated and consequently accelerating research.

Consolidation of predictions. Results from different approaches were not entirely integrated. In structure-based repurposing approaches, candidate drugs obtained from different docking tools or homology modeling methods could be consolidated to provide an ensemble of repurposable drugs (Fig. 3b). For host-targeting therapies, one study used rank aggregation to integrate results from different algorithms⁶⁹. Another study derived the final predictions by combining the output of their model with results from gene set enrichment and expert knowledge⁶⁸. While it should be noted that the drugs in clinical trials were used to develop the methods, these two studies predicted the highest proportion of overlaps with drugs being tested in clinical trials. The latter shows the potential of ensemble approaches, which are well known to output more robust results^{91,92}. Consolidation of multiple approaches could significantly increase confidence for repurposing candidates and guide clinical researchers through the drug selection process. This requires a streamlined solution, considering tool accessibility and standardization, as in a standardized database that stores drug candidate predictions enabling meta-analyses.

Combinatorial treatment development. Computationally identifying synergistic drug combinations is an underexplored domain which could provide highly valuable information to augment clinical decision-making, since they have been demonstrated to be more effective than finding monotherapies^{91,92} (Fig. 3c). So far, targeting of viral and host proteins has been performed independently. There is a lack of methods aiming to find complementary drug groups while considering side effects. Combining drugs from both virus- and host-targeting categories is a promising strategy that acts by blocking the viral and host molecular machinery required for SARS-CoV-2 entry into cells and disrupting the host pathways involved in disease progression in combination with inhibitors for

viral replication. While thousands of compounds can be evaluated *in vitro*⁹⁰, combinatorial validations are considerably more challenging. Predicted combinatorial treatments could drastically reduce the search space for subsequent *in vitro* validation. Existing screening databases such as the NIH OpenData portal⁹³ or the ReFRAME library⁹⁴ have been sparsely used, but their potential has not been exhausted. By extending them with *in silico* predictions, they could link *in silico* and *in vitro* research, and help identify promising combinatorial treatments. Furthermore, screening results help verify computational predictions. Especially for docking simulations, model predictions and parameters can be easily released in a standardized format, which can be evaluated by experimental researchers. For host-targeting therapies, the study of Zhou et al.⁶⁸ is an example of a combinatorial approach. Furthermore, several trials are registered for combination therapy that include candidate drugs from both categories; of these, ten drugs were included in the predictions from the reviewed studies (Supplementary Table 10). However, these drugs are either in the recruitment phase or limited results were reported; thus, data regarding their effectiveness has been inconclusive.

Expert knowledge. Limited understanding of the complex biological mechanisms underlying COVID-19 has required expert knowledge or manual curation in certain stages of the workflow, either at protein or pathway selection or at filtering of drug predictions (Fig. 3d). Expert vetting is mainly intended to uncover inconsistent or contradictory results while still allowing the identification of new predictions and can be crucial for filtering candidate drug lists for possible adverse side effects. To illustrate this, the antimalarial drug (hydroxy)chloroquine raised concerns regarding its potential toxicity. Chlorhexidine was found by a docking-based study⁴⁰ as a potential drug targeting SARS-CoV-2 proteins; however, chlorhexidine is a widely used disinfectant whose mechanism of action is not SARS-CoV-2-specific and it is approved for topical or dental application only⁹⁵. Consequently, the use of expert knowledge for careful evaluation of potential repurposable drugs would have been helpful to allocate limited experimental and computational resources on safe and effective drugs that have greater potential for widespread application. Close collaboration between computational and clinical researchers is therefore crucial, because computational approaches are still limited in side effect data and annotations for drug actions on the targets.

Validation strategies. Drug repurposing studies usually validate the computational models by constructing their own ‘ground truth’; these can include data from *in vitro* screening of predicted compounds, *in vivo* experiments using animal models, ongoing clinical trials, electronic health records, literature mining or expert knowledge⁹⁶ (Fig. 3e). Thus, there is considerable heterogeneity in the sources of these standards, but efforts are ongoing to address this. For instance, newly released databases, such as the NIH’s OpenData portal⁹³, collect and continuously update SARS-CoV-2 *in vitro* screening data for thousands of compounds and other SARS-CoV-2-related assays. We encourage future studies to utilize such resources for further validation or filtering of *in silico* predictions. However, except for one study,⁶⁹ no direct follow-up experimental validation has been performed in the drug repurposing efforts for COVID-19. In the reviewed studies, validation was implemented through several strategies. Some studies performed signature matching of drug profiles or gene set enrichment analysis¹⁷ to provide evidence of the potential effectiveness^{69,72}. Others evaluated the performance of their pipelines using the drugs undergoing clinical trials for COVID-19^{17,69} or experimental results from *in vitro* drug screening⁶⁹. However, an extensive list of candidate drugs remains experimentally invalidated; thus, systematic validation of candidate drugs would be required to provide a landscape of the accuracy of

methods. Since this is infeasible in practice, combining the predictions with expert knowledge becomes even more important.

The proposed strategy in this work has the potential to address the gaps of previous studies and is intended to serve as a guideline on computational drug repurposing to accelerate research, promote standardization, and react faster and more precisely in the case of future pandemics.

Received: 16 September 2020; Accepted: 1 December 2020;

Published online: 14 January 2021

References

1. Pushpakom, S. et al. Drug repurposing: progress, challenges and recommendations. *Nat. Rev. Drug Discov.* **18**, 41–58 (2019).
2. Paranjpe, M. D., Taubes, A. & Sirota, M. Insights into computational drug repurposing for neurodegenerative disease. *Trends Pharmacol. Sci.* **40**, 565–576 (2019).
3. Sanseau, P. & Koehler, J. Computational methods for drug repurposing. *Brief. Bioinform.* **12**, 301–302 (2011).
4. Ciliberto, G. & Cardone, L. Boosting the arsenal against COVID-19 through computational drug repurposing. *Drug Discov. Today* **25**, 946–948 (2020).
5. Benson, D. A. et al. GenBank. *Nucleic Acids Res.* **41**, D36–D42 (2013).
6. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data – from vision to reality. *Euro Surveill.* **22**, (2017).
7. The UniProt Consortium. The Universal Protein Resource (UniProt). *Nucleic Acids Res.* **35**, D193–D197 (2007).
8. Berman, H. M. The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
9. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
10. Duan, Q. et al. LINCS Canvas Browser: interactive web app to query, browse and interrogate LINCS L1000 gene expression signatures. *Nucleic Acids Res.* **42**, W449–W460 (2014).
11. Gordon, D. E. et al. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* **583**, 459–468 (2020).
12. Navratil, V. et al. VirHostNet: a knowledge base for the management and the analysis of proteome-wide virus-host interaction networks. *Nucleic Acids Res.* **37**, D661–D668 (2009).
13. Guirimand, T., Delmotte, S. & Navratil, V. VirHostNet 2.0: surfing on the web of virus/host molecular interactions data. *Nucleic Acids Res.* **43**, D583–D587 (2015).
14. Kotlyar, M., Pastrello, C., Sheahan, N. & Jurisica, I. Integrated interactions database: tissue-specific view of the human and model organism interactomes. *Nucleic Acids Res.* **44**, D536–D541 (2016).
15. Szklarczyk, D. et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).
16. Percha, B. & Altman, R. B. A global network of biomedical relationships derived from text. *Bioinformatics* **34**, 2614–2624 (2018).
17. Zeng, X. et al. Repurpose open data to discover therapeutics for COVID-19 using deep learning. *J. Proteome Res.* **19**, 4624–4636 (2020).
18. Wishart, D. S. et al. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* **36**, D901–D906 (2008).
19. Gaulton, A. et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **40**, D1100–D1107 (2012).
20. Sterling, T. & Irwin, J. J. ZINC 15 – ligand discovery for everyone. *J. Chem. Inf. Model.* **55**, 2324–2337 (2015).
21. Yoshino, R., Yasuo, N. & Sekijima, M. Identification of key interactions between SARS-CoV-2 main protease and inhibitor drug candidates. *Sci. Rep.* **10**, 12493 (2020).
22. Al-Khafaji, K., Al-Duhaidahawi, D. & Taskin Tok, T. Using integrated computational approaches to identify safe and rapid treatment for SARS-CoV-2. *J. Biomol. Struct. Dyn.* <https://doi.org/10.1080/07391102.2020.1764392> (2020).
23. Alamri, M. A. et al. Pharmacoinformatics and molecular dynamic simulation studies reveal potential inhibitors of SARS-CoV-2 main protease 3CL^{pro}. *J. Biomol. Struct. Dyn.* <https://doi.org/10.1080/07391102.2020.1782768> (2020).
24. Arya, R., Das, A., Prashar, V. & Kumar, M. Potential inhibitors against papain-like protease of novel coronavirus (SARS-CoV-2) from FDA approved drugs. Preprint at <https://doi.org/10.26434/chemrxiv.11860011.v2> (2020).
25. Chang, Y.-C. et al. Potential therapeutic agents for COVID-19 based on the analysis of protease and RNA polymerase docking. Preprint at <https://doi.org/10.20944/preprints202002.0242.v1> (2020).
26. Chen, Y. W., Yiu, C.-P. B. & Wong, K.-Y. Prediction of the SARS-CoV-2 (2019-nCoV) 3C-like protease (3CL^{pro}) structure: virtual screening reveals velpatasvir, ledipasvir, and other drug repurposing candidates. *F1000Res.* **9**, 129 (2020).

27. Elfiky, A. A. Ribavirin, Remdesivir, Sofosbuvir, Galidesivir, and Tenofovir against SARS-CoV-2 RNA dependent RNA polymerase (RdRp): A molecular docking study. *Life Sci.* **253**, 117592 (2020).
28. Elfiky, A. & Ibrahim, N. S. Anti-SARS and anti-HCV drugs repurposing against the Papain-like protease of the newly emerged coronavirus (2019-nCoV). Preprint at <https://doi.org/10.21203/rs.2.23280/v1> (2020).
29. Gao, K., Nguyen, D. D., Wang, R. & Wei, G.-W. Machine intelligence design of 2019-nCoV drugs. Preprint at <https://doi.org/10.1101/2020.01.30.927889> (2020).
30. Gupta, M. K. et al. In-silico approaches to detect inhibitors of the human severe acute respiratory syndrome coronavirus envelope protein ion channel. *J. Biomol. Struct. Dyn.* <https://doi.org/10.1080/07391102.2020.1751300> (2020).
31. Hall, D. C. & Ji, H.-F. A search for medications to treat COVID-19 via in silico molecular docking models of the SARS-CoV-2 spike glycoprotein and 3CL protease. *Travel Med. Infect. Dis.* **35**, 101646 (2020).
32. Hosseini, F. S. & Amanlou, M. Simeprevir, potential candidate to repurpose for coronavirus infection: virtual screening and molecular docking study. *Life Sci.* **258**, 118205 (2020).
33. Khan, R. J. et al. Targeting SARS-CoV-2: a systematic drug repurposing approach to identify promising inhibitors against 3C-like proteinase and 2'-O-ribose methyltransferase. *J. Biomol. Struct. Dyn.* <https://doi.org/10.1080/07391102.2020.1753577> (2020).
34. Khan, S. A., Zia, K., Ashraf, S., Uddin, R. & Ul-Haq, Z. Identification of chymotrypsin-like protease inhibitors of SARS-CoV-2 via integrated computational approach. *J. Biomol. Struct. Dyn.* <https://doi.org/10.1080/07391102.2020.1751298> (2020).
35. Li, Y. et al. Therapeutic drugs targeting 2019-nCoV main protease by high-throughput screening. Preprint at <https://doi.org/10.1101/2020.01.28.922922> (2020).
36. Lin, S., Shen, R., He, J., Li, X. & Guo, X. Molecular modeling evaluation of the binding effect of ritonavir, lopinavir and darunavir to severe acute respiratory syndrome coronavirus 2 proteases. Preprint at <https://doi.org/10.1101/2020.01.31.929695> (2020).
37. Muralidharan, N., Sakthivel, R., Velmurugan, D. & Gromiha, M. M. Computational studies of drug repurposing and synergism of lopinavir, oseltamivir and ritonavir binding with SARS-CoV-2 protease against COVID-19. *J. Biomol. Struct. Dyn.* <https://doi.org/10.1080/07391102.2020.1752802> (2020).
38. Smith, M. & Smith, J. C. Repurposing therapeutics for COVID-19: supercomputer-based docking to the SARS-CoV-2 viral spike protein and viral spike protein-human ACE2 interface. Preprint at <https://doi.org/10.26434/chemrxiv.11871402.v4> (2020).
39. Wang, J. Fast identification of possible drug treatment of coronavirus disease-19 (COVID-19) through computational drug repurposing Study. *J. Chem. Inf. Model.* **60**, 3277–3286 (2020).
40. Wu, C. et al. Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods. *Acta Pharm. Sin. B* **10**, 766–788 (2020).
41. Xu, Z. et al. Nelfinavir was predicted to be a potential inhibitor of 2019-nCoV main protease by an integrative approach combining homology modelling, molecular docking and binding free energy calculation. Preprint at <https://doi.org/10.1101/2020.01.27.921627> (2020).
42. Ton, A.-T., Gentile, F., Hsing, M., Ban, F. & Cherkasov, A. Rapid identification of potential inhibitors of SARS-CoV-2 main protease by deep docking of 1.3 billion compounds. *Mol. Inform.* **39**, e2000028 (2020).
43. Talluri, S. Virtual high throughput screening based prediction of potential drugs for COVID-19. Preprint at <https://doi.org/10.20944/preprints202002.0418.v1> (2020).
44. Beck, B. R., Shin, B., Choi, Y., Park, S. & Kang, K. Predicting commercially available antiviral drugs that may act on the novel coronavirus (SARS-CoV-2) through a drug-target interaction deep learning model. *Comput. Struct. Biotechnol. J.* **18**, 784–790 (2020).
45. Forli, S. et al. Computational protein-ligand docking and virtual drug screening with the AutoDock suite. *Nat. Protoc.* **11**, 905–919 (2016).
46. Halgren, T. A. et al. Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem.* **47**, 1750–1759 (2004).
47. Gentile, F. et al. Deep Docking: a deep learning platform for augmentation of structure based drug discovery. *ACS Cent. Sci.* **6**, 939–949 (2020).
48. Torres, P. H. M., Sodero, A. C. R., Jofily, P. & Silva, F. P. Jr Key topics in molecular docking for drug design. *Int. J. Mol. Sci.* **20**, 4574 (2019).
49. Nguyen, D. D., Gao, K., Wang, M. & Wei, G.-W. MathDL: mathematical deep learning for D3R Grand Challenge 4. *J. Comput. Aided Mol. Des.* **34**, 131–147 (2020).
50. Nguyen, D. D., Gao, K., Chen, J., Wang, R. & Wei, G.-W. Potentially highly potent drugs for 2019-nCoV. Preprint at <https://doi.org/10.1101/2020.02.05.936013> (2020).
51. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**, 31–36 (1988).
52. Lee, S. M.-Y. & Yen, H.-L. Targeting the host or the virus: current and novel concepts for antiviral approaches against influenza virus infection. *Antiviral Res.* **96**, 391–404 (2012).
53. Min, J.-Y. & Subbarao, K. Cellular targets for influenza drugs. *Nat. Biotechnol.* **28**, 239–240 (2010).
54. Catanzaro, M. et al. Immune response in COVID-19: addressing a pharmacological challenge by targeting pathways triggered by SARS-CoV-2. *Signal Transduct. Target. Ther.* **5**, 84 (2020).
55. Liao, J., Way, G. & Madahar, V. Target virus or target ourselves for COVID-19 drugs discovery?—Lessons learned from anti-influenza virus therapies. *Medi. Drug Discov.* **5**, 100037 (2020).
56. Chen, L. et al. Clinical characteristics of pregnant women with Covid-19 in Wuhan, China. *N. Engl. J. Med.* **382**, e100 (2020).
57. Lamb, J. et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**, 1929–1935 (2006).
58. Subramanian, A. et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* **171**, 1437–1452.e17 (2017).
59. Loganathan, T., Ramachandran, S., Shankaran, P., Nagarajan, D. & Mohan, S. S. Host transcriptome-guided drug repurposing for COVID-19 treatment: a meta-analysis based approach. *PeerJ* **8**, e9357 (2020).
60. Jia, Z., Song, X., Shi, J., Wang, W. & He, K. Transcriptome-based drug repositioning for coronavirus disease 2019 (COVID-19). *Pathog. Dis.* **78**, rta036 (2020).
61. Law, J. N. et al. Identifying human interactors of SARS-CoV-2 proteins and drug targets for COVID-19 using network-based label propagation. Preprint at <https://arxiv.org/abs/2006.01968> (2020).
62. Hoffmann, M. et al. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *Cell* **181**, 271–280.e8 (2020).
63. Messina, F. et al. COVID-19: viral–host interactome analyzed by network based-approach model to study pathogenesis of SARS-CoV-2 infection. *J. Transl. Med.* **18**, 233 (2020).
64. Sadegh, S. et al. Exploring the SARS-CoV-2 virus–host–drug interactome for drug repurposing. *Nat. Commun.* **11**, 3518 (2020).
65. Gyöngyi, Z., Garcia-Molina, H. & Pedersen, J. Combating web spam with TrustRank. In *Proc. 2004 VLDB Conference* (eds Nascimento, M. A. et al.) 576–587 (Morgan Kaufmann, 2004).
66. Brin, S. & Page, L. The anatomy of a large-scale hypertextual web search engine. **30**, 107–117 (1998).
67. Cheng, F. et al. Network-based approach to prediction and population-based validation of in silico drug repurposing. *Nat. Commun.* **9**, 2691 (2018).
68. Zhou, Y. et al. Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. *Cell Discov* **6**, 14 (2020).
69. Gysi, D. M. et al. Network medicine framework for identifying drug repurposing opportunities for COVID-19. Preprint at <https://arxiv.org/abs/2004.07229> (2020).
70. List, M. et al. KeyPathwayMinerWeb: online multi-omics network enrichment. *Nucleic Acids Res.* **44**, W98–W104 (2016).
71. Belyaeva, A. et al. Causal network models of SARS-CoV-2 expression and aging to identify candidates for drug repurposing. Preprint at <https://arxiv.org/abs/2006.03735> (2020).
72. Ge, Y. et al. A data-driven drug repositioning framework discovered a potential therapeutic agent targeting COVID-19. Preprint at <https://doi.org/10.1101/2020.03.11.986836> (2020).
73. Favipiravir plus hydroxychloroquine and lopinavir/ritonavir plus hydroxychloroquine in COVID-19. *ClinicalTrials.gov* <https://clinicaltrials.gov/ct2/show/NCT04376814> (2020).
74. Baricitinib therapy in COVID-19. *ClinicalTrials.gov* <https://clinicaltrials.gov/ct2/show/NCT04358614> (2020).
75. Lopinavir/ritonavir, ribavirin and IFN-beta combination for nCoV treatment. *ClinicalTrials.gov* <https://clinicaltrials.gov/ct2/show/NCT04276688> (2020).
76. An investigation into beneficial effects of interferon beta 1a, compared to interferon beta 1b and the base therapeutic regimen in moderate to severe COVID-19: a randomized clinical trial. *ClinicalTrials.gov* <https://clinicaltrials.gov/ct2/show/NCT04343768> (2020).
77. Cao, B. et al. A trial of lopinavir–ritonavir in adults hospitalized with severe Covid-19. *N. Engl. J. Med.* **382**, 1787–1799 (2020).
78. Lopinavir-Ritonavir results. *RECOVERY trial* (2020); <https://www.recoverytrial.net/results/lopinavar-results>
79. 'Solidarity' clinical trial for COVID-19 treatments. *WHO* (accessed November 2020); <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/global-research-on-novel-coronavirus-2019-ncov/solidarity-clinical-trial-for-covid-19-treatments>
80. Trial of treatments for COVID-19 in hospitalized adults. *ClinicalTrials.gov* (2020); <https://clinicaltrials.gov/ct2/show/NCT04315948>

81. Beigel, J. H. et al. Remdesivir for the treatment of Covid-19. *N. Engl. J. Med.* **383**, 1813–1826 (2020).
82. Grein, J. et al. Compassionate use of remdesivir for patients with severe Covid-19. *N. Engl. J. Med.* **382**, 2327–2336 (2020).
83. Goldman, J. D. et al. Remdesivir for 5 or 10 days in patients with severe Covid-19. *N. Engl. J. Med.* **383**, 1827–1837 (2020).
84. Wang, M. et al. Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-nCoV) in vitro. *Cell Res.* **30**, 269–271 (2020).
85. Huang, N., Shoichet, B. K. & Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **49**, 6789–6801 (2006).
86. Wang, R., Fang, X., Lu, Y. & Wang, S. The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J. Med. Chem.* **47**, 2977–2980 (2004).
87. The RECOVERY Collaborative Group. Dexamethasone in hospitalized patients with Covid-19. *New Engl. J. Med.* <https://doi.org/10.1056/nejmoa2021436> (2020).
88. WHO Rapid Evidence Appraisal for COVID-19 Therapies (REACT) Working Group. et al. Association between administration of systemic corticosteroids and mortality among critically ill patients with COVID-19: a meta-analysis. *JAMA* **324**, 1330–1341 (2020).
89. Zhang, Y., Xu, Q., Sun, Z. & Zhou, L. Current targeted therapeutics against COVID-19: Based on first-line experience in China. *Pharmacol. Res.* **157**, 104854 (2020).
90. Jin, Z. et al. Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors. *Nature* **582**, 289–293 (2020).
91. Sun, W., Sanderson, P. E. & Zheng, W. Drug combination therapy increases successful drug repositioning. *Drug Discov. Today* **21**, 1189–1195 (2016).
92. Liu, H. et al. Predicting effective drug combinations using gradient tree boosting based on features extracted from drug-protein heterogeneous network. *BMC Bioinform.* **20**, 645 (2019).
93. Brimacombe, K. R. et al. An OpenData portal to share COVID-19 drug repurposing data in real time. Preprint at <https://doi.org/10.1101/2020.06.04.135046> (2020).
94. Janes, J. et al. The ReFRAME library as a comprehensive drug repurposing library and its application to the treatment of cryptosporidiosis. *Proc. Natl Acad. Sci. USA* **115**, 10750–10755 (2018).
95. Syed Shihaab, S. & Pradeep Chlorhexidine: its properties and effects. *Res. J. Pharm. Technol.* **9**, 1755–1760 (2016).
96. Jarada, T. N., Rokne, J. G. & Alhaji, R. A review of computational drug repositioning: strategies, approaches, opportunities, challenges, and directions. *J. Cheminform.* **12**, 46 (2020).

Acknowledgements

J.B. was partially funded by his VILLUM young investigator grant no. 13154. M.S.A. received PhD fellowship funding from CONACYT (CVU659273) and the German Academic Exchange Service, DAAD (ref. 91693321). This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement nos 777111 and 826078. This publication reflects the authors' views only and the European Commission is not responsible for any use that may be made of the information it contains. This project is funded by the Bavarian State Ministry of Science and the Arts in the framework of the Bavarian Research Institute for Digital Transformation (bidit).

Author contributions

G.G., J.M., T.D.R., S.S., M.S.A., J.S., J.B. and J.K.P. contributed equally to the manuscript writing. J.B. and J.K.P. were in charge of overall direction, planning and supervision. All authors provided critical feedback and helped to improve the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s43588-020-00007-6>.

Correspondence should be addressed to J.K.P.

Peer review information Fernando Chirigati was the primary editor on this Review and managed its editorial process and peer review in collaboration with the rest of the editorial team. *Nature Computational Science* thanks Arnab Chatterjee, Brian Shoichet, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

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

© Springer Nature America, Inc. 2021


A.4. Lacking mechanistic disease definitions and corresponding association data hamper progress in network medicine and beyond

Lacking mechanistic disease definitions and corresponding association data hamper progress in network medicine and beyond

Received: 29 July 2022

Accepted: 13 March 2023

Published online: 25 March 2023

 Check for updates

Sepideh Sadegh ^{1,2}, James Skelton³, Elisa Anastasi³, Andreas Maier ², Klaudia Adamowicz², Anna Möller ⁴, Nils M. Kriege ^{5,6}, Jaanika Kronberg ⁷, Toomas Haller⁷, Tim Kacprowski ^{8,9}, Anil Wipat^{3,11}, Jan Baumbach ^{2,10,11} & David B. Blumenthal ^{4,11} 

A long-term objective of network medicine is to replace our current, mainly phenotype-based disease definitions by subtypes of health conditions corresponding to distinct pathomechanisms. For this, molecular and health data are modeled as networks and are mined for pathomechanisms. However, many such studies rely on large-scale disease association data where diseases are annotated using the very phenotype-based disease definitions the network medicine field aims to overcome. This raises the question to which extent the biases mechanistically inadequate disease annotations introduce in disease association data distort the results of studies which use such data for pathomechanism mining. We address this question using global- and local-scale analyses of networks constructed from disease association data of various types. Our results indicate that large-scale disease association data should be used with care for pathomechanism mining and that analyses of such data should be accompanied by close-up analyses of molecular data for well-characterized patient cohorts.

Since the seminal articles by Goh et al.¹ and Barabási et al.², network medicine has developed into an increasingly mature and diverse research field with its own dedicated journals³, associations⁴, and subfields. One of the network medicine field's long-term objectives is to replace our current mainly phenotype-based disease classification systems by a mechanistically grounded disease vocabulary^{5–7}. In such a vocabulary, phenotype-based disease definitions are replaced by so-called endotypes, i.e., distinct molecular mechanisms underlying the disease phenotypes. Once properly disentangled into disjoint,

individually targetable endotypes⁵, disease-modifying treatment strategies might become available for diseases which, at the moment, can be treated only symptomatically.

Two clarifications are required to define the scope of this paper: Firstly, we use the term “endotype” to denote molecular endotypes as explained by Anderson⁸, Lötvalld et al.⁹, and Nogales et al.⁵ – i.e., the underlying molecular mechanisms driving disease phenotypes. There are other works where the term “endo(patho)phenotype” denotes common intermediate phenotypes⁶ such as inflammation, fibrosis, or

¹Chair of Experimental Bioinformatics, TUM School of Life Sciences, Technical University of Munich, Munich, Germany. ²Institute for Computational Systems Biology, University of Hamburg, Hamburg, Germany. ³School of Computing, Newcastle University, Newcastle upon Tyne, UK. ⁴Biomedical Network Science Lab, Department Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany. ⁵Faculty of Computer Science, University of Vienna, Vienna, Austria. ⁶Research Network Data Science, University of Vienna, Vienna, Austria. ⁷Estonian Genome Centre, Institute of Genomics, University of Tartu, Tartu, Estonia. ⁸Division Data Science in Biomedicine, Peter L. Reichertz Institute for Medical Informatics of Technische Universität Braunschweig and Hannover Medical School, Braunschweig, Germany. ⁹Braunschweig Integrated Centre of Systems Biology (BRICS), TU Braunschweig, Braunschweig, Germany. ¹⁰Computational Biomedicine Lab, Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark. ¹¹These authors jointly supervised this work: Anil Wipat, Jan Baumbach, David B. Blumenthal. ✉ e-mail: david.b.blumenthal@fau.de

thrombosis which drive phenotypic disease manifestations^{10,11}. Secondly, we would like to stress that compiling an endotype-based disease vocabulary is a genuinely biomedical rather than a semantic endeavor: It does not consist in redefining semantic relationships between existing disease terms but in uncovering currently unknown molecular disease mechanisms and dissecting umbrella diseases such as Alzheimer's disease or coronary artery disease into endotypes which are clearly characterized at a molecular level⁵.

In order to reach the objective of an endotype-based disease vocabulary, network medicine approaches aim at uncovering pathomechanisms driving diseases. Here, we broadly distinguish between close-up and bird's-eye-view (BEV) network medicine approaches, depending on the data used as primary input towards this task (this distinction is of course an idealized binarization of a continuous spectrum, but serves as a conceptual framework for this article). Close-up network medicine studies focus on a specific disease and start their analyses with molecular data for well-characterized patient cohorts. Such studies are typically carried out as close collaborations between bioinformaticians and domain experts from the biomedical sciences. They tend to be time- and labor-intensive and often involve the development or customization of data analysis methods for specific datasets. The most impressive translational results of the network medicine field have been reached via such close-up studies. For instance, close-up studies have led to novel mechanistic insights into type 2 diabetes¹², liver fibrosis¹³, pulmonary arterial hypertension¹⁴, asthma¹⁵, hypertrophic cardiomyopathy¹⁶, pre-eclampsia¹⁷, chronic obstructive pulmonary disease, and idiopathic pulmonary fibrosis¹⁸.

In contrast to that, BEV approaches use large-scale disease association data that are typically gathered from several data sources. Various studies have generated evidence for the validity of this overall approach: For instance, Menche et al.¹⁹ demonstrated that disease-associated genes form so-called disease modules, i.e., highly connected subnetworks within protein-protein interaction (PPI) networks, and that biological and clinical similarity of two diseases results in significant topological proximity of these modules. In a similar vein, Iida et al.²⁰ showed that shared therapeutic targets or shared drug indications are correlated with high topological module proximity. Guney et al.²¹ and Cheng et al.²² showed that the network-based separation between drug targets and disease modules is indicative of drug efficacy. Cheng et al.²³ and Zhou et al.²⁴ found that FDA-approved drug combinations are proximal to each other and to the modules of the targeted diseases in the interactome.

Despite the promising findings summarized above, several studies have pointed out important biases in the data used by BEV approaches. Menche et al.¹⁹ have studied the effect of incompleteness of disease-gene association and protein-protein interaction (PPI) data on network medicine. Schaefer et al.²⁵ have shown that the previously observed^{26–28} high node degree of cancer-associated proteins in PPI networks can largely be explained by the fact that cancer-associated proteins are tested more often for interaction than others. Lazareva et al.²⁹ found that widely used methods to mine PPI networks for pathomechanisms inherit this bias in that they mainly learn from the node degrees instead of exploiting the biological knowledge encoded in the edges of the PPI networks. Haynes et al.³⁰ showed that study bias also distorts functional gene annotation resources such as the Gene Ontology (GO)³¹. Kustatcher et al.³² made a similar point for functional protein annotations and sketched a roadmap for systematically exploring the understudied part of the proteome. Stoeger et al.³³ and Rodriguez-Esteban³⁴ looked into reasons that might lead to the emergence of gene study bias and identified, respectively, a limited number of biological characteristics³³ and speed of information propagation between scientific communities as potential drivers³⁴.

While the aforementioned studies have analyzed the impact of various types of data biases related to genes and proteins (and, to a lesser extent, also variants), the disease part of disease-gene and other

disease association data introduces another, so far unstudied type of data bias: In currently available large-scale disease association data, diseases are annotated with the very phenotype-based disease definitions the network medicine field aims to overcome. BEV approaches hence risk to systematically reproduce the biases introduced by these disease definitions. Consequently, BEV approaches make the implicit assumption that the biases introduced by phenotype-based disease definitions even out and that, despite those biases, disease association data using these definitions still contain useful information about the pathomechanism that are to be uncovered.

In this work, we quantify to which extent this implicit assumption is indeed backed by data. Towards this end, we construct disease-disease networks (called “diseasomes” in the remainder of this article) based on (1) disease-gene associations, (2) disease-variant associations, (3) comorbidity data, (4) symptom data, and (5) drug-indication data, as well as drug-disease and drug-drug networks (called “drugomes”) based on drug-indication and drug-target data. We then formulate two testable hypotheses that follow from the implicit assumption of BEV network medicine: The global-scale hypothesis states that, globally, networks constructed from two different types of association data that both contain useful information about endotypes should be pairwise more similar than expected by chance. The local-scale hypothesis states that this should hold not only globally but also for the neighborhoods of the individual diseases and drugs represented by nodes in the constructed networks.

In line with the findings of prior studies^{20–24}, our analyses provide solid evidence for the global-scale hypothesis. However, they only partially support the local-scale hypothesis. Figuratively speaking, BEV network medicine hence only allows a distal view at the endotypes that are to be discovered. When zooming in on individual diseases, the picture becomes blurred and less reliable (see Fig. 1 for a conceptual visualization and Fig. 2 for a concrete exemplification of this phenomenon in the context of neurodegenerative diseases). This implies that, in order to yield translational results, BEV approaches need to be supplemented with additional layers of molecular data for well-characterized patient cohorts and a dedicated focus on the specific diseases which are being investigated. In particular, fine-grained molecular patient data are crucial for implementing network medicine's long-term objective to replace current phenotype- or organ-based disease definitions by mechanistically grounded endotypes. The main finding of this study is hence that the biases current disease definitions introduce in large-scale disease association databases such as OMIM and DisGeNET do not even out and that such databases should be used with care in all fields of data-centric biomedicine: Instead of blindly using public disease association data out of convenience for pathomechanism mining, we strongly recommend biomedical researchers to always consciously ponder to which extent biases in these data introduced by phenotype-based disease terms threaten to distort their potential findings.

Results

Neurodegenerative diseases as case example

Before presenting the comprehensive results of our analyses, we visualize the phenomenon of local blurriness in BEV network medicine with a small example. We compiled a list of diseases that fall under the parent term “neurodegenerative disease” in the MONDO disease hierarchy. From those, we kept diseases for which we have nodes in the aligned gene- and drug-based diseasomes. This led to a cluster of seven neurodegenerative diseases which are highly connected in both diseasomes. Figure 2 shows this cluster, together with the contained diseases' local empirical *P*-values obtained from the comparison of gene- and drug-based diseasomes in MONDO space, the global empirical *P*-value, as well as the cluster-level empirical *P*-value (see next subsection and Methods for explanations on how we obtained the *P*-values). While only two local empirical *P*-values are significant at 0.05

level, the cluster-level and global empirical P -values are significant at levels 0.01 and 0.001, respectively.

Overview of analyses

Let D be disease association data of some data type T commonly used by BEV approaches (e.g., disease-gene associations). Further assume that D contains entries $D(d_1)$ and $D(d_2)$ for two diseases d_1 and d_2 that

share an unknown molecular disease mechanism. Then this shared mechanism should lead to similarities between $D(d_1)$ and $D(d_2)$, given that D indeed contains useful information about disease mechanisms³⁵. For instance, we would expect that the diseases d_1 and d_2 have similar profiles of disease-associated genes, that they exhibit high comorbidity, that they lead to similar symptoms, and that they can be treated by similar drugs. We can capture such similarities in

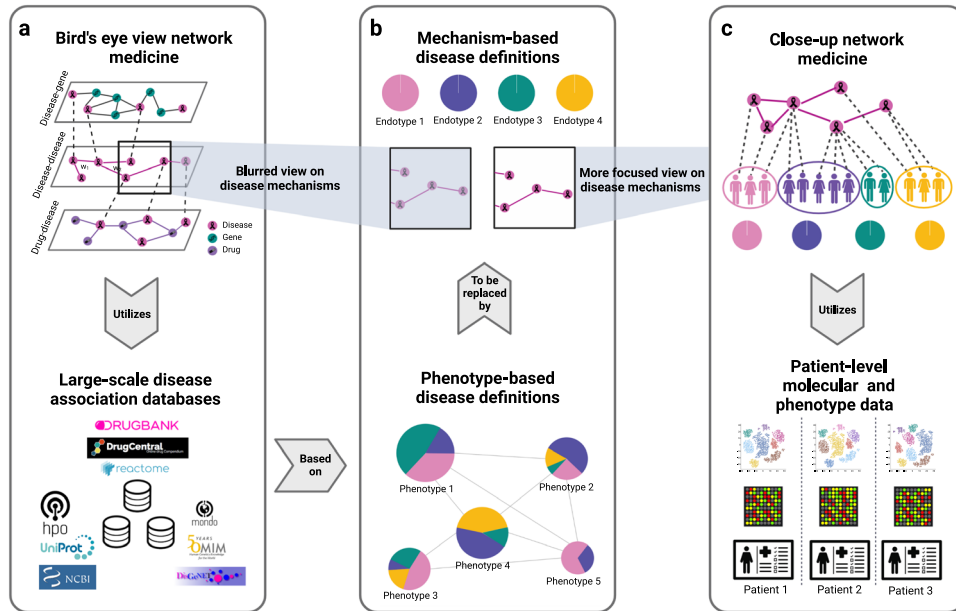


Fig. 1 | BEV vs. close-up network medicine. **a** BEV network medicine mainly utilizes large-scale disease association data where diseases are annotated with phenotype-based disease definitions (**b**, bottom). BEV network medicine inherits the bias introduced by these definitions, which leads to a blurred view on individual

pathomechanisms (**b**, top). **c** Close-up network medicine uses patient-level molecular data and is hence less dependent on the phenotype-based disease definitions that network medicine aims to replace by mechanism-based endotypes.

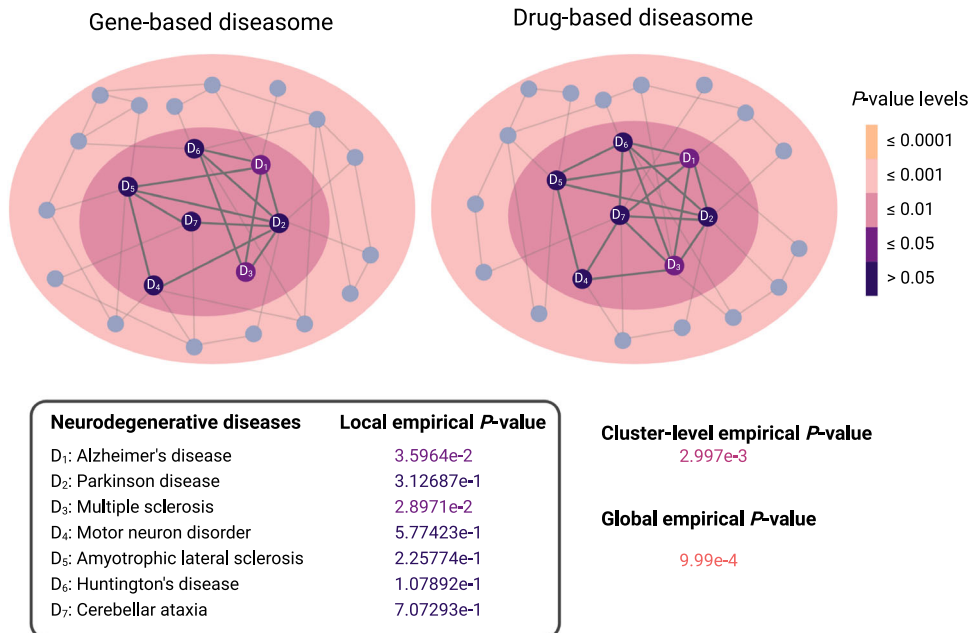


Fig. 2 | Locally blurred results for neurodegenerative diseases. The color gradient visualizes local-, global-, and cluster-level empirical P -values (one-sided, unadjusted) obtained from the comparison of gene- and drug-based diseaseomes in MONDO vocabulary. The gene-based diseaseome was constructed based on disease-gene association data integrated from DisGeNET³⁶ and OMIM¹³ and two diseases

were connected by an edge if they share at least one disease associated gene. The drug-based diseaseome was constructed based on drug-indication data integrated from CTD⁴⁸ and DrugCentral³⁷ and two diseases were connected by an edge if they share at least one indicated drug.

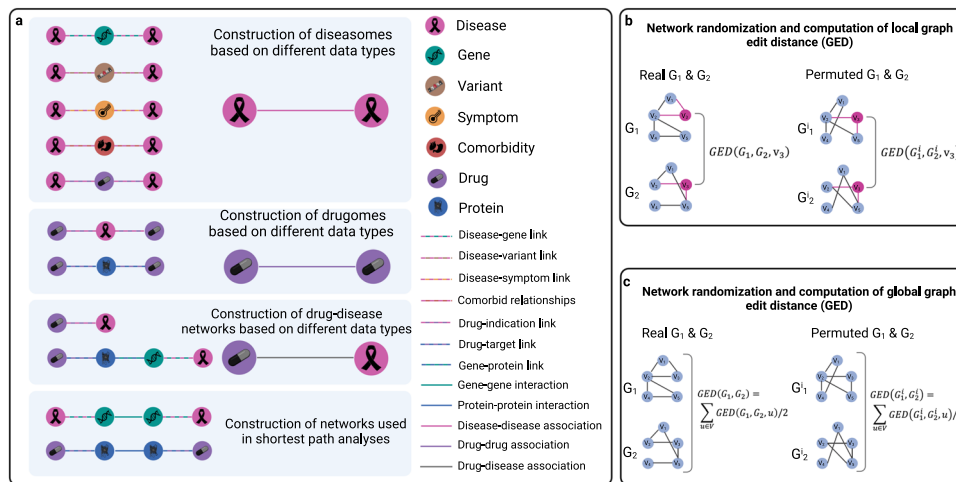


Fig. 3 | Overview of compared networks and graph edit distance computation.

a We compared five different types of disease-disease networks (diseaseomes), two different types of drug-drug networks (drugomes), and two different types of drug-disease networks. Pairwise comparisons between those networks were carried out using local and global graph edit distance (GED). **b** Local GED was used to quantify

the dissimilarities of the individual nodes' neighborhoods across different networks in comparison to pairs of randomly rewired networks. **c** Global network dissimilarities were computed using global GED, obtained by summing up the local GEDs of the individual nodes.

diseaseomes, where diseases d_1 and d_2 are connected by an edge if $D(d_1)$ and $D(d_2)$ are sufficiently similar. In order to assess the implicit assumption of BEV network medicine approaches with quantitative means, we hence formulate the following testable hypotheses (see Methods for an argument to support these hypotheses):

- Global-scale hypothesis: For all disease association data D_1 and D_2 that are assumed to contain useful information about endotypes (e.g., disease-gene association and drug-indication data from databases such as DisGeNET³⁶ and DrugCentral³⁷), diseaseomes G_1 and G_2 constructed based on D_1 and D_2 should be pairwise more similar than expected by chance.
- Local-scale hypothesis: For all disease association data D_1 and D_2 that are assumed to contain useful information about endotypes and any disease term d that appear in D_1 and D_2 , the direct neighborhood of d in the diseaseomes G_1 and G_2 constructed based on D_1 and D_2 should be pairwise more similar than expected by chance. For example, under the assumption that disease-gene and drug-indication databases such as DisGeNET and DrugCentral contain useful information about Alzheimer's disease (AD) mechanisms, there should be a significant overlap between the set of diseases whose associated genes overlap with AD-associated genes and the set of diseases which can be treated with drugs also indicated for AD.

To test these two hypotheses, we constructed various diseaseomes, drugomes, and drug-disease networks based on different data types. An overview of the used data types and derived networks is shown in Fig. 3a. Using customized versions of the graph edit distance (GED)^{38,39}, we then compared these networks in a pairwise manner both on a local scale, i.e. zoomed-in on individual disease or drug nodes, and on a global scale. More precisely, we generated 1000 permuted networks as randomized counterparts for each network. Subsequently, we compared the distributions of local and global GEDs obtained for the original networks to GED distributions obtained for randomized counterparts. Network randomization and computation of local and global GED are illustrated in Fig. 3b, c. While local GED measures the dissimilarity between the individual nodes' neighborhoods in the compared networks, global GED is a measure for the overall dissimilarity of the networks.

We also evaluated how annotating the data using disease vocabularies of different granularity affect the results, by carrying out the

analyses using MONDO IDs⁴⁰ and UMLS CUIs⁴¹ (finer granularity) and ICD-10⁴² three-character codes (coarser granularity) as node IDs in the constructed networks, respectively. To this end, where possible, we constructed the networks in MONDO, UMLS CUI, and in ICD-10 vocabulary (using three-character level codes). Note that analyses involving comorbidity data were carried out only in ICD-10 and the comparison between target- and indication-based drugomes only in MONDO vocabulary (see Methods for an explanation). Moreover, neither the semantic layers of the MONDO disease ontology nor the hierarchy of the UMLS CUI and ICD-10 classification system were used to add edges to our diseaseomes. MONDO, UMLS CUI, and ICD-10 were only used as vocabularies, i.e., to provide the node IDs in our networks. Whether two disease nodes are connected by an edge exclusively depends on the primary databases containing the association data (upon mapping to MONDO, UMLS CUI, or ICD-10). For instance, two diseases are connected in the gene-based diseaseome in MONDO vocabulary if the intersection of the sets of genes associated with their MONDO IDs is non-empty, where disease-gene associations were obtained from OMIM⁴³ and DisGeNET³⁶.

GED quantifies the dissimilarity between two networks as the minimum cost of an edit path transforming one network into the other. Edit paths are sequences of elementary edit operations (node and edge insertions, substitutions, and deletions), all of which come with associated edit costs. Hence, the GED is a distance measure between two networks. We computed three different versions of GED using uniform, weight-based, and rank-based edge editing costs, respectively. Uniform edit costs discard the association strengths of the edges in the compared networks; weight- and rank-based edit costs incorporate them by making it more expensive to delete or insert edges with strong associations or to substitute them by edges with weak associations. Corroborating the robustness of our analysis method, we obtained similar results for all three versions of GED. In the following, only the results of uniform edit costs are reported. Results for rank- and weight-based edit costs can be found in Supplementary Figs. 1–4 and 9–12, respectively. More details on disease vocabulary mapping, network construction, and GED computation can be found in Methods.

Results of global-scale analyses

To test the global-scale hypothesis, we computed empirical P -values for each pair of networks based on global GEDs (Fig. 4a, left panel). For all evaluated pairs of networks (in MONDO, UMLS CUI, and ICD-10

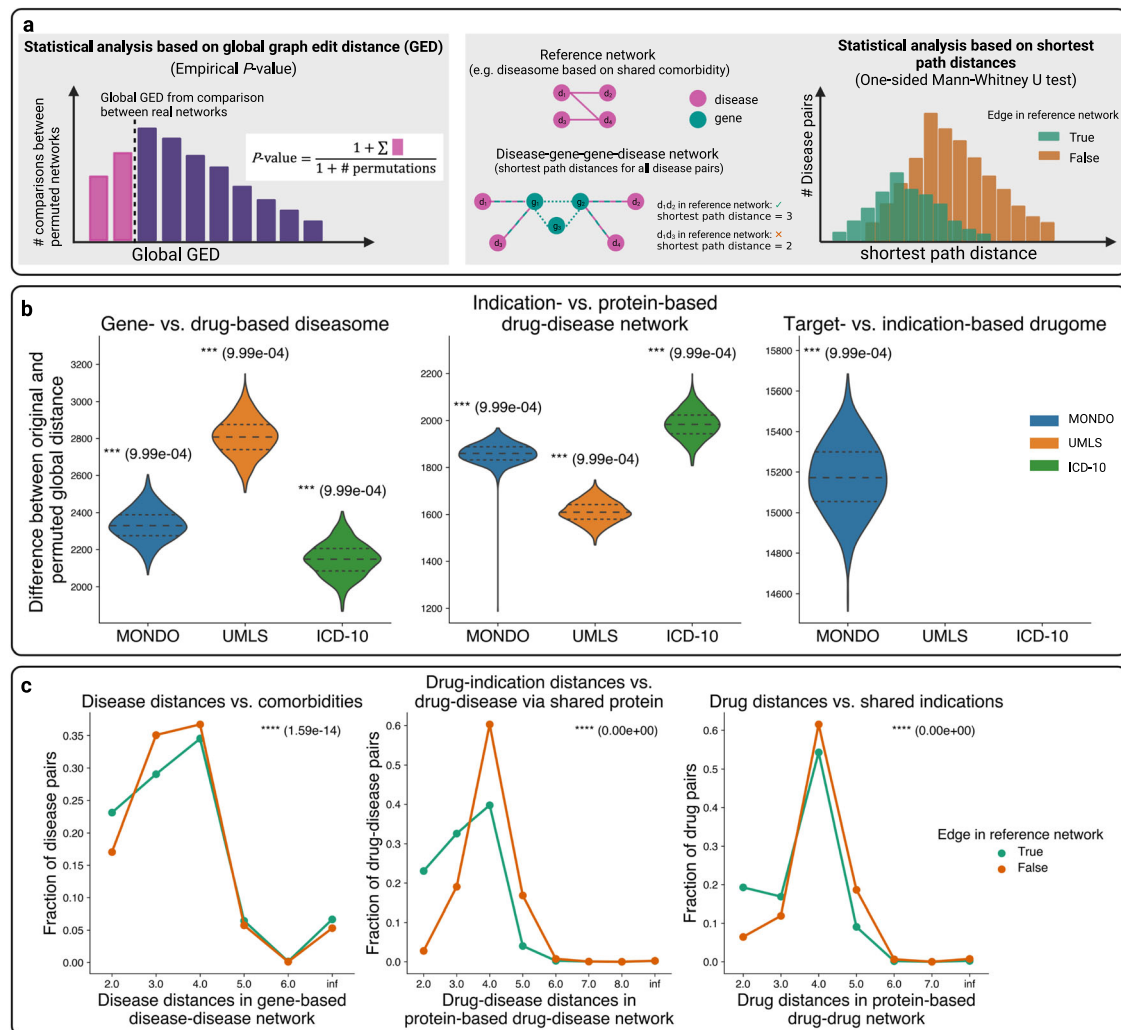


Fig. 4 | Global-scale analyses. **a** Illustration of global-scale analysis methods. Left panel: Statistical analyses based on global GED via empirical P -values. Right panel: Statistical analyses based on shortest path distances via MWU test. **b** Differences of global GEDs (based on uniform edge edit costs) between a selection of original networks and their counterpart permuted networks, and corresponding global empirical P -values (one-sided, unadjusted) in MONDO, UMLS, and ICD-10 vocabularies. All obtained global empirical P -values are at the lower resolution limit of our permutation tests with 1000 randomized network pairs. **c** Selected results of

shortest path analyses and the corresponding MWU P -values (one-sided, unadjusted). Left: Disease distances in gene-based disease-disease network vs. comorbidity-based diseaseome as the reference network. Middle: Drug-disease distances in protein-based drug-disease network vs. drug-indication network as the reference network. Right: Drug distances in protein-based drug-drug network vs. indication-based drugome as the reference network. All networks underlying the results shown in (c) are constructed in the MONDO vocabulary.

vocabularies), we obtained smaller global GEDs for the original diseaseomes, drugomes, or drug-disease networks than for randomized counterparts, leading to empirical P -values which are significant at 0.001 level. Differences between GEDs obtained for permuted and a selection of original networks are shown in Fig. 4b. For the full results of our global-scale analyses, see Supplementary Fig. 5.

Moreover, we performed analyses based on shortest path distances between disease-disease, drug-drug, and drug-disease pairs in disease-gene-gene-disease, drug-protein-protein-drug, and disease-protein-protein-drug networks, where protein-protein and gene-gene links were obtained from PPIs. We then compared shortest path distances for node pairs which do and node pairs which do not have a link in different reference networks, using the Mann-Whitney U (MWU) test (Fig. 4a, right panel).

For all shortest path analyses, we observed that shortest path distances are significantly shorter for node pairs that are connected by a link in the reference networks (see Fig. 4c for a selection of the results). In particular, the results show (1) that distances between diseases that are connected by edges in diseaseomes constructed based on

comorbidities, shared drugs, shared symptoms, or shared genetic variants are significantly shorter than distances between diseases without such edges (Supplementary Fig. 6a–d); (2) that distances of disease-drug pairs with shared indication edges are significantly shorter than distances of disease-drug pairs without such edges (Supplementary Fig. 6e); and (3) that distances between drug pairs with shared indication are significantly shorter than distances for drug pairs without shared indications (Supplementary Fig. 6f). In sum, our global analyses hence provide solid evidence for the global validity of the BEV network medicine paradigm and hence further corroborate the findings of previous studies^{19–24}.

Results of local-scale analyses

To test the local-scale hypothesis, we computed P -values using the one-sided MWU test based on local GEDs to evaluate whether the local distances for the original networks are significantly smaller than the local distances for the permuted counterparts (Fig. 5a, left panel). Local GEDs of nodes obtained for the permuted and a selection of original networks and the corresponding MWU P -values are shown in

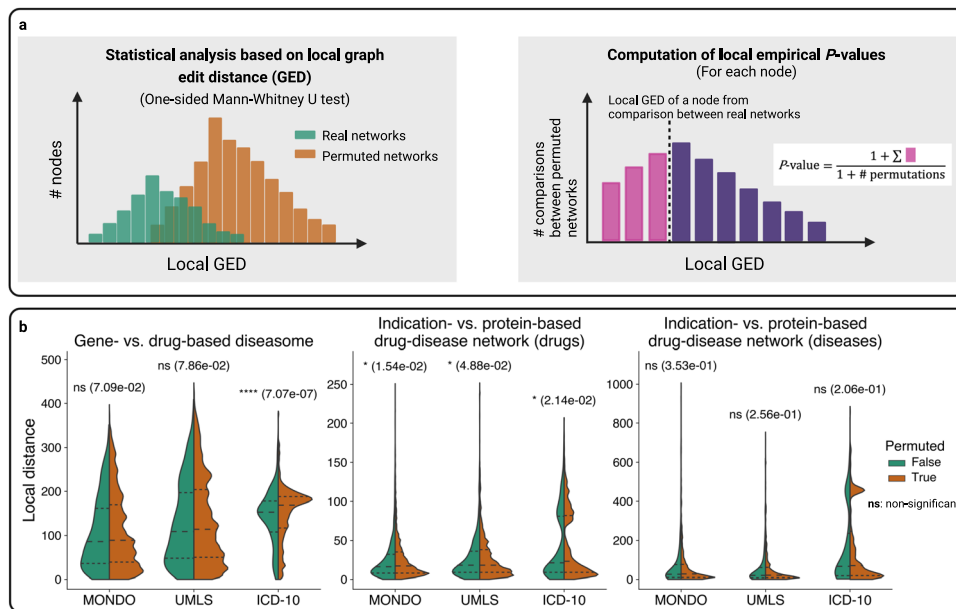


Fig. 5 | Local-scale analyses: methods and local GEDs. **a** Illustration of local-scale analysis methods. Left panel: Statistical analyses based on local GED via MWU test. Right panel: Computation of empirical P -values (one-sided, unadjusted) of each node based on local GEDs. **b** Local GEDs (of all nodes) between a selection of original networks vs. their permuted counterpart networks and corresponding

MWU P -values. Left: Similarities between gene- and drug-based diseases. Middle: Similarities between indication- and protein-based drug-disease network (for drugs). Right: Similarities between indication- and protein-based drug-disease network (for diseases). Results shown in **(b)** are based on uniform edge edit cost.

Fig. 5b (for the full results of the local-scale analyses, see Supplementary Fig. 7). The overview of the results of the local GED analyses in different vocabularies shows that the comparisons performed in ICD-10 vocabulary (at three-character level) led to more significant similarities than the ones performed in MONDO or UMLS CUI vocabulary (Fig. 6a and Supplementary Fig. 4a). As an example, the P -value computed from the local GEDs of drug-based vs. gene-based diseases in ICD-10 vocabulary is significant at 0.0001 level ($P \approx 7.1 \times 10^{-7}$), while it is not significant in the MONDO and UMLS CUI vocabularies ($P \approx 0.071$ for MONDO, $P \approx 0.079$ for UMLS CUI).

The results of the MWU test for local GED analyses point out that we have more significant similarities in ICD-10 (8 out of 10 significant at 0.05 level) than in MONDO vocabulary (2 out of 6 significant at 0.05 level) or UMLS CUI vocabulary (1 out of 6 significant at 0.05 level). The results also suggest that variant-based diseases have higher similarities with other diseases (7 out of 10 comparisons significant at 0.05 level) than gene-based diseases (5 out of 10 comparisons significant at 0.05 level), considering all three vocabularies. By inspecting the P -values of drug nodes (3 out of 3 comparisons significant at 0.05 level) against disease nodes (0 out of 3 comparisons significant at 0.05 level) obtained from local-similarity analyses of indication- versus protein-based drug-disease network as well as P -values obtained from target- and indication-based drugome (significant at 0.001 level), we discovered that, in general, drug neighborhoods are better preserved across the compared networks than disease neighborhoods (Fig. 6a, bottom right panel).

Furthermore, we computed local empirical P -values individually for nodes based on local GEDs (Fig. 5a, right panel). The local empirical P -values for all network comparisons are shown in Supplementary Fig. 8. The fractions of significant local empirical P -values at 0.05 level are shown in Fig. 6b and Supplementary Figs. 4b and 12b. Our results show that, for a substantial fraction of disease nodes, local neighborhoods are preserved not only not significantly better but worse than expected by chance across the different diseases (compare sigmoidal shape of curves in Supplementary Fig. 8). The local-scale hypothesis hence seems to hold for some diseases, but does not hold at all for others.

In follow-up analyses, we tried to identify patterns explaining these results, e.g., by assessing whether there are certain chapters of the ICD-10 disease vocabulary which are enriched with diseases with very small or very large empirical P -values. However, no clear patterns could be discovered, indicating that it is very hard to predict for which concrete diseases BEV network medicine approaches can be expected to yield robust and reliable results. Our local analyses hence only provide weak evidence for the local-scale hypothesis, indicating the BEV network medicine tends to produce locally blurred results.

Web tool for interactive exploration of results

In order to make our results explorable and actionable, we developed the GraphSimViz (graph similarity visualizer) web interface, which is freely available at <https://graphsimgviz.net>. GraphSimViz allows biomedical researchers to query and visualize our findings for user-selected drugs, diseases, network types, and disease vocabularies. Using GraphSimViz, biomedical researchers can assess if a specific type of disease association data is likely to contain reliable information about pathomechanisms underlying their diseases of interest. Below, we illustrate how GraphSimViz can be employed for interactive exploration of our results, using neurodegenerative diseases as a case example. To enable quantification of the effect of biases introduced by mechanistically ungrounded disease definitions in data sources not covered by our study, we provide the GraphSimQT (graph similarity quantification tool) Python package, which is freely available on GitHub (<https://github.com/repotial/graphsimqt>).

Discussion

Our results strongly support the global-scale hypothesis and, in line with previous studies^{19–24}, provide solid evidence for the overall validity of the BEV network medicine paradigm. However, they also indicate that results generated via BEV network medicine approaches become less reliable when zooming-in on individual diseases. Our results hence confirm that it is problematic to exclusively rely on data annotated with phenotype-based definitions if the objective is to uncover molecular pathomechanisms. As long as phenotype-based disease definitions have not been replaced by endotypes,

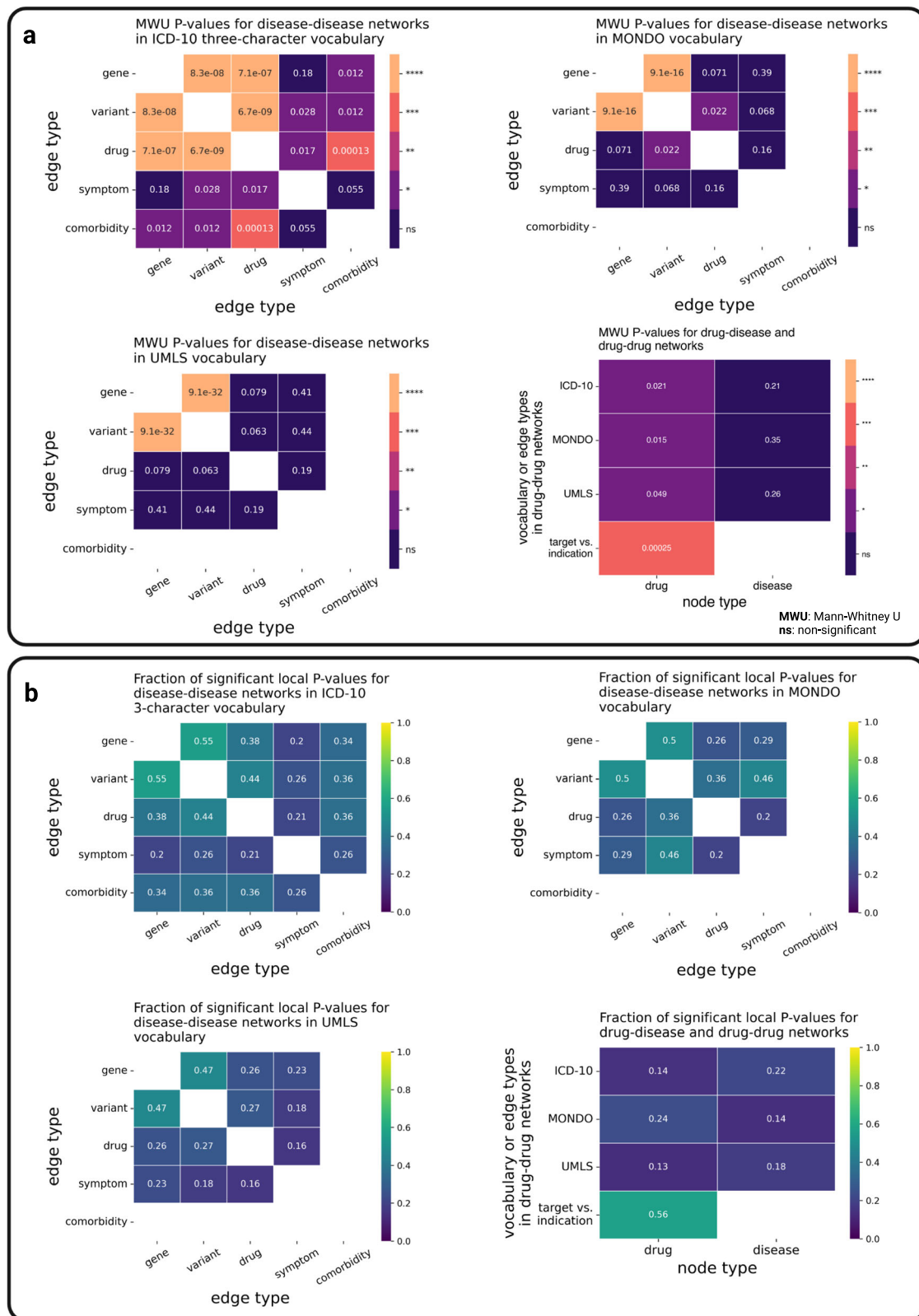


Fig. 6 | Local-scale analyses: MWU P-values and local empirical P-values. **a** Overview of MWU P-values (one-sided, unadjusted) computed from local GEDs with levels of significance. **b** Fraction of significant local empirical P-values (one-

sided, unadjusted) at 0.05 level computed from local GEDs on a pair of networks for the original vs. permuted network. All results are based on uniform edge edit cost.

large-scale disease association databases should therefore be used with care in network medicine and should be combined with additional layers of disease-specific omics data. In the following, we further speculate on issues that might play a role in the local

blurriness of BEV network medicine and sketch a roadmap to overcome this problem.

While there are vast amounts of datasets online that contain useful information about diseases such as genetic associations,

comorbidities, and symptoms, each of these datasets may use different disease vocabularies to describe their associations. The vocabularies have different degrees of granularity and are generated in different ways and for different purposes. However, for downstream (BEV) network medicine analyses, in order to jointly leverage the disease association from various data sources that use disease terms from different vocabularies as disease identifiers, we have to map data to a joint target vocabulary. This is a mammoth task that inevitably involves losing some data due to unmappable terms (see Fig. 7 for the levels of completeness of disease vocabulary mappings underlying this study).

The choice of the disease vocabulary has the potential to dramatically affect the results of downstream analyses (see discordant results of local-scale analyses carried out using ICD-10 three-character codes, on the one hand, and UMLS CUIs or MONDO IDs, on the other hand, shown in Fig. 6a and Supplementary Fig. 4a). At the same time, for most analysis tasks, the choice of the disease vocabulary is dictated by the format of the data and, thus, often impossible to change without losing information at the time of analysis. The vocabularies used to annotate disease-associated data must hence be viewed as confounders which are very difficult if not impossible to control for.

Currently used disease vocabularies are not only used discordantly, but also mechanistically inadequate: Since causal molecular disease mechanisms are often unknown, disease names often do not

denote such mechanisms but rather reflect the person who coined the disease term (e.g., “Alzheimer’s disease”), areas in the body that are affected (e.g., “kidney stones”) or symptoms of the disease (e.g., “irritable bowel syndrome”). ICD-10 codes are considered inadequate due to their overly inclusive designations, ranging from symptoms (e.g., cough) over syndromes (e.g., cachexia) to true endotypes with definable molecular determinants (e.g., Mendelian disorders). This leads to data that is blurred, as diseases with distinct pathomechanisms are being aggregated together, e.g., due to symptom or organ commonality. This blurriness not only has severe clinical consequences (patients with mechanistically distinct diseases receive the same untargeted treatment), but also makes it very challenging to mine disease-associated data for pathomechanisms via BEV network medicine approaches⁴⁴. Since such analyses often require case-versus-control or subtype annotations as input, it is very difficult to obtain meaningful results if the employed disease definitions are too unspecific.

The results presented in this study, where drugome comparisons have led to more significant results on a local level than diseaseome comparisons, are evidence that network-based analyses yield more targeted and reliable results when the underlying annotations are well-defined (such as in drug vocabularies). Comparing the results of the GED-based analyses for full diseaseomes (global analyses) with those obtained for analyses based on local GEDs in diseaseomes with ICD-10 three-character codes, UMLS CUIs, and MONDO terms as nodes, respectively, further highlights the detrimental effect of local blurriness in currently used disease definitions: The higher the resolution of the analysis, the less significant the obtained *P*-values (see Fig. 8). When using MONDO or UMLS CUI terms (fine granularity) as nodes in the diseaseomes, only the comparisons between gene- and variant-based diseaseomes consistently (with respect to uniform, weight-based, and rank-based edit costs) led to smaller local distances in the original networks than in their randomized counterparts. No other network comparisons in the MONDO or UMLS vocabularies yielded significant *P*-values for all three types of edit costs. When using ICD-10 three-character codes (which denote disease clusters rather than individual diseases), around 50% of all computed MWU *P*-values are significant at 0.001 level. When comparing the entire diseaseomes via global GEDs, all empirical *P*-values are significant.

The fact that we could not identify any clear patterns among diseases with small or large empirical *P*-values computed based on local GEDs may be a consequence of some of the current phenotype-based disease entities already corresponding to true endotypes. We speculate that, for diseases where our current definitions already have a one-to-one mapping to true endotypes, the local-scale hypothesis holds.

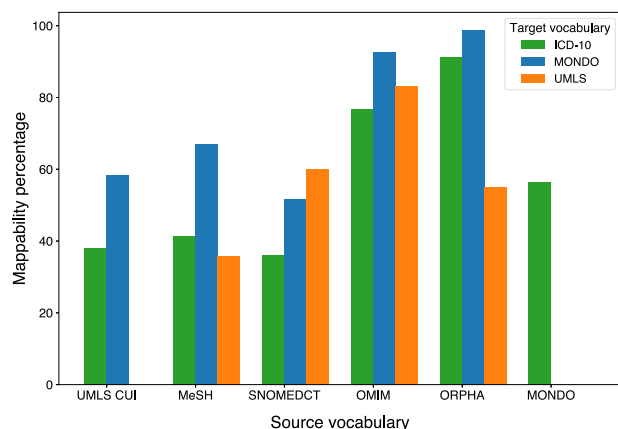


Fig. 7 | Levels of completeness of disease vocabulary mappings underlying this article. For each source-target vocabulary pair, mappability is computed as the percentage of terms in the source vocabulary used in this study that could be mapped to a term in the target vocabulary.

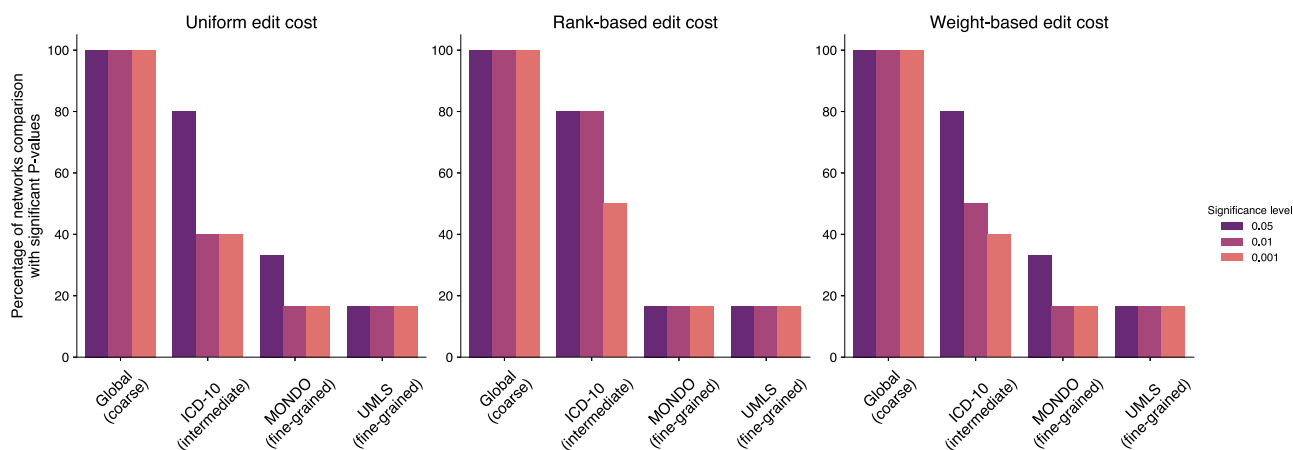


Fig. 8 | Effect of disease term granularity on results of GED-based analyses. For the individual *P*-values summarized in this figure, see Fig. 6a, as well as Supplementary Figs. 1, 4a, 5, 9, and 12a.

Even though we expected to obtain similar results for variant-based and gene-based diseases, the local-similarity analyses show that variant-based diseases have higher similarities with other diseases compared to gene-based diseases. This indicates that the disease-gene associations underlying the gene-based diseases contain less targeted information than the disease-variant associations underlying the variant-based diseases. Hence, using disease-variant data might yield more reliable results in the context of BEV network medicine applications.

To seek a possible explanation for this difference, we had a closer look at the associations underlying these two types of diseases. In our study, as well as in many other network medicine studies^{1,2,22,45,46}, disease-gene associations were taken from OMIM and DisGeNET curated databases. The latter collates disease-gene associations from different databases: UniProt⁴⁷, CTD⁴⁸, Orphanet⁴⁹, ClinGen⁵⁰, Genomics England⁵¹, CGI⁵², and PsyGeNET⁵³. These constituent databases comprise multiple types of disease-gene associations such as causal mutations (mutations known to cause the disease), modifying mutations (mutations known to modify the clinical presentation of the disease), or merely statistical associations without evidence of causality. Disease-variant associations used in our study were extracted from DisGeNET, which itself integrates various databases: GWASdb⁵⁴, ClinVar⁵⁵, GWAS Catalog⁵⁶, UniProt, and BeFree⁵⁷. Like for disease-gene associations, there are different types of disease-variant associations, ranging from known causal variants to variants with merely statistical evidence. However, the heterogeneity of the association types is higher for disease-gene associations than for disease-variant associations. Moreover, the genetic variation data from the constituent disease-variant databases of DisGeNET is mainly taken from genome-wide association studies (GWAS), which identify associations between common genetic variants and phenotypic traits via hypothesis-free, genome-wide scans. In contrast, in the disease-gene databases used by DisGeNET, parts of the data are curated from studies where evidence for disease-gene associations stems from a very limited number of patients or where hypothesis-driven approaches were used (i.e. the analyzed genetic variants were limited to those contained in candidate genes selected a priori).

Another reason for the difference in results between gene-based and variant-based diseases may consist in the loss of detail resulting from mapping variants to genes. Distinct mutations in one gene may cause different phenotypes, but this information cannot be captured at the level of disease-gene associations and is better conserved at disease-variant level. A very good example is the LMNA gene, where different mutations can cause 13 different diseases such as Hutchinson-Gilford progeria syndrome and the Dunnigan-type familial partial lipodystrophy⁵⁸. Finally, the difference in results between gene- and variant-based diseases may also partly be due to loss of information introduced when aggregating *P*-values for disease-variant associations at gene level⁵⁹.

A limitation of our study is that our results do not rule out the possibility that confounders other than mechanistically inadequate disease definitions lead to the observed local blurriness of BEV network medicine. For instance, off-target effects might introduce biases in our analyses using drug association data, while the known biases in gene association data discussed above might explain the results obtained for analyses involving gene association data. However, we would like to stress that the obtained results are remarkably stable across all employed data modalities (see distributions of the obtained local empirical *P*-values in Supplementary Figs. 3, 8, and 11). Since phenotype-based disease definitions are the only confounders that affect all data types, this is strong (but of course not conclusive) evidence that the observed local blurriness can indeed mainly be attributed to them.

We started our investigation with the question of whether biases introduced by phenotype- and organ-based disease mechanisms even

out when mining large-scale disease association data for disease mechanisms – an assumption implicitly made by BEV medical research approaches. Our results indicate that this question has to be answered negatively, which has several consequences for the network medicine field and beyond.

Firstly, our findings imply that uncritical use of databases such as DisGeNET or OMIM which rely on phenotype-based disease definitions is problematic. Instead, we emphasize that close-up approaches remain the gold standard in network medicine, where data scientists collaborate with researchers from the biomedical sciences and jointly analyze molecular as well as deep phenotype data for the same patients. In such a collaborative setup, a positive feedback loop can emerge, where initial hypotheses about disease subtypes and their underlying pathomechanisms are formulated based on the analysis of molecular data, further refined using deep phenotyping (e.g., histological images, blood-derived biomarkers, etc.) and expert knowledge of the clinicians, and finally validated in preclinical studies (e.g., gain- or loss-of-function studies). As mentioned above, such approaches have already led to various important insights into specific disease mechanisms.

Secondly, unsupervised network medicine methods are needed, which not only return candidate pathomechanisms but at the same time de novo stratify patients into mechanistically distinct subgroups and hence do not rely on potentially misleading priorly available phenotypically defined subtype annotations. While few such approaches exist^{60–62}, most existing pathomechanism mining methods still rely on phenotypic case-versus-control annotations^{63,64} or lists of genes associated with a (potentially ill-defined) disease term^{65–67}.

Finally, we would like to point out that the current lack of mechanistic disease definitions not only hampers progress in (BEV) network medicine, but also has a detrimental effect on virtually all other data-centric approaches to, e.g., treatment design or diagnosis which rely on disease association data that utilize phenotype-based disease definitions. For instance, an artificial intelligence model for diagnosis assistance trained on genetic disease signatures will systematically produce unreliable results if the disease annotations used for training do not correspond to true endotypes. While we here quantified the effect of this problem in the context of BEV medicine, overcoming it would hence be beneficial for a large fraction of the biomedical research community.

Methods

Compliance with ethical regulations

Our research complies with all relevant ethical regulations. The only non-public data used for this study is the comorbidity data we obtained from the Estonian Biobank. The Estonian Biobank is a population-based biobank managed by the Institute of Genomics at the University of Tartu. All participants have signed a broad consent upon joining the biobank, allowing their sample and data to be used for further research. ICD-10 diagnoses are obtained from epicrisis, prescriptions and bills to the Health Insurance Fund. The work in this article was covered by the ethics approval “234T-12 Omics for Health” (March 19, 2014) by the Estonian Committee of Bioethics and Human Research. Data was released by the Estonian Biobank (release M11, July 24, 2019).

Data integration

As shown in Table 1, the data sources used to create the different networks use a range of competing disease vocabularies to refer to diseases. We hence had to map these vocabularies to a common vocabulary to be able to investigate network (dis-)similarities. The similarity analyses were performed in MONDO (Monarch Disease Ontology), UMLS CUI, and ICD-10 vocabularies. Disease ID mapping to MONDO and ICD-10 was carried out via the two-step approach implemented in the NeDRex platform⁶⁸. First, MONDO contains mappings between its own disease vocabulary and various other

Table 1 | Data sources used for network construction

Data source	Used disease vocabularies	Data type	Networks constructed from data source
HPO ⁵⁶	OMIM, Orphanet (ORPHA)	Disease-symptom	Symptom-based diseasome
DisGeNET	Concept Unique Identifiers of Unified Medical Language System (UMLS CUI)	Disease-gene, disease-variant	Gene-based diseasome, variant-based diseasome, disease-gene-gene-disease network, drug-protein-protein-drug network, drug-protein-protein-disease network
OMIM	OMIM	Disease-gene	Gene-based diseasome, disease-gene-gene-disease network, drug-protein-protein-disease network
DrugCentral ³⁷	SNOMED Clinical Terms ⁸⁷ (SNOMEDCT)	Drug-target, drug-indication	Target-based drugome, indication-based drugome and drug-disease network, drug-protein-protein-drug network, drug-protein-protein-disease network
DrugBank ⁸⁸	-	Drug-target	Target-based drugome, drug-protein-protein-drug network, drug-protein-protein-disease network
CTD ⁴⁸	MeSH	Drug-indication	Drug-disease network, indication-based drugome
IID ⁸⁹	-	Protein-protein interaction	Disease-gene-gene-disease network, drug-protein-protein-drug network, drug-protein-protein-disease network
UniProt	-	Gene-protein	Drug-protein-protein-disease network
Estonian Biobank ⁹⁰	ICD-10 (mixed three- and four-character codes)	Comorbidity data	Comorbidity-based diseasome

vocabularies, including OMIM, MeSH⁶⁹, and ICD-10. Then, mappings between several vocabularies and ICD-10 could be achieved by mapping disease terms to MONDO, followed by mapping MONDO to ICD-10. Mapping to UMLS CUI was carried out using the mappings provided in the UMLS Metathesaurus 2022AA full release. For all pairwise analyses, the two compared networks were aligned before computing GEDs, i.e., only the nodes contained in both of them were taken into account.

The comorbidity data was obtained from the Estonian Biobank, which uses originally ICD-10 codes. In order to carry out analyses involving comorbidity data in MONDO or UMLS CUI vocabulary, the comorbidity data needed to be mapped from a coarser-grained (ICD-10) to a finer-grained disease vocabulary (MONDO and UMLS CUI). Although this is possible from a technical point of view, it would have introduced a lot of noise in the obtained comorbidity networks. To avoid overshadowing all other effects by the introduced noise, we decided to carry out analyses involving comorbidity data only in ICD-10 vocabulary. Consequently, all analyses involving comorbidity data were carried out only in ICD-10 vocabulary. On the other hand, the comparison between the target- and the indication-based drugomes was carried out only in MONDO vocabulary. In these networks, nodes are drugs and not diseases and using different disease vocabularies leaves the nodes of the networks unchanged. In the indication-based drugomes, the choice of the disease vocabulary can change the edges of the networks, but, in practice, we observed that the differences are small. Target-based drugomes are not affected at all by the choice of the disease ontology. Therefore, we only use MONDO for the comparison of drugomes.

Additionally, further data harmonization steps were carried out: Since HPO contains both general and specific terms, we pruned the data by removing very general symptom terms, using the existing hierarchy in HPO. More specifically, we decomposed the generated hierarchical phenotype network into its levels and removed the terms from the top three levels.

The diagnoses in around 140 K patients records available in the Estonian Biobank (April 2020 version used for this study) are encoded in ICD-10 vocabulary, and the records contain both three- and four-character ICD-10 codes. In order to generate uniform data, we therefore truncated all four-character codes to three-character level. Moreover, we removed diseases with incidence below five from the data, as well as the codes from the ICD-10 chapters XV (“Pregnancy, childbirth and the puerperium”), XVI (“Certain conditions originating in the perinatal period”), XVIII (“Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified”), XIX (“Injury,

poisoning and certain other consequences of external causes”), XX (“External causes of morbidity and mortality”), XXI (“Factors influencing health status and contact with health services”), and XXII (“Codes for special purposes”).

Network construction

For network construction, some part of the data such as disease-gene, drug-indication, drug-target, gene-encoding-protein, and PPI data were obtained from the databases shown in Table 1, using the data access and mapping provided by the NeDRex platform⁶⁸. Disease-variant and disease-symptom associations were directly obtained from DisGeNET and HPO, respectively.

Supplementary Table 1 shows the most important properties of all constructed networks. The comorbidity-based diseasome was constructed via ϕ -correlation. Let I_i denote the incidence of disease i and C_{ij} be the number of patients who were simultaneously diagnosed with diseases i and j . The comorbidity between the two diseases can be measured by

$$\phi_{ij} = \frac{C_{ij}N - I_i I_j}{\sqrt{I_i I_j (N - I_i)(N - I_j)}}, \quad (1)$$

where N is the total number of patient records ($N=139,065$ for the Estonian Biobank data). When two diseases co-occur more frequently than expected by chance, we have $\phi_{ij}>0$. We used one-tailed Fisher’s exact test followed by Benjamini-Hochberg correction for multiple testing to determine the significance of comorbidity associations and connected two diseases by an edge if adjusted $P \leq 0.05$. Edge weights were defined using the ϕ -correlation, i.e., we set $w_{ij} = \phi_{ij}$ for all diseases i and j with significant comorbidity association.

The indication- and target-based drugomes as well as the gene-, variant-, symptom-, and indication-based diseasomes were constructed based on the Jaccard index of the respective annotations. A_i denotes the set of annotations for a disease or drug i used as node in the network under construction (e.g., when constructing the gene-based diseasome, A_i is the set of all genes associated with disease i). We connected diseases i and j by an edge if $|A_i \cap A_j| \geq 1$ and defined the edge weights as $w_{ij} = |A_i \cap A_j| / |A_i \cup A_j|$. Disease nodes with $|A_i| = 0$ were removed from the networks, i.e., empty annotation sets were treated as missing data.

The bipartite indication-based drug-disease network was directly constructed from the data source, i.e., we connected a disease i with a drug j if i is an indication for j . For the bipartite target-based drug-

disease network, we connected a disease i with a drug j if j targets a protein encoded by a gene associated to i . In both drug-disease networks, edges are unweighted. Finally, we constructed drug-protein-protein-disease networks where drugs are connected to their targets, experimentally validated PPIs from IID are used to connect proteins, and diseases are connected to proteins encoded by disease-associated genes.

Graph edit distance

GED is a widely used and generically applicable distance measure for attributed graphs^{38,39,70}. It is defined as the minimum cost of transforming a source graph $G_1 = (V_1, E_1)$ into a target graph $G_2 = (V_2, E_2)$ via elementary edit operations, i.e., by deleting, inserting, and substituting nodes and edges. Equivalently, GED can be defined as the minimum edit cost induced by a node map π from G_1 to G_2 , where nodes maps $\pi \subseteq (V_1 \cup \{\epsilon_1\}) \times (V_2 \cup \{\epsilon_2\})$ are relations that cover all nodes $u \in V_1$ and $v \in V_2$ exactly once (ϵ_1 and ϵ_2 are dummy nodes that may be covered multiple times or left uncovered)⁷¹.

We used a customized version of GED to compare the different diseasesomes, drugomes, and drug-disease networks constructed as detailed in the previous section as well as their randomized counterparts. Since the networks were aligned before all pairwise comparisons, we had $V_1 = V_2 = V$ (node sets are identical) whenever comparing two networks. Consequently, we fixed π as the identity and computed GED as the sum of edge edit costs induced by the identity (the edge edit cost functions sub, del, and ins are explained below):

$$GED(G_1, G_2) = \sum_{uv \in E_1 \cap E_2} \text{sub}(uv) + \sum_{uv \in E_1 \setminus E_2} \text{del}(uv) + \sum_{uv \in E_2 \setminus E_1} \text{ins}(uv) \quad (2)$$

$GED(G_1, G_2)$ quantifies the global distance between the graphs G_1 and G_2 . Since the node sets of G_1 and G_2 are identical in our analyses, it can be decomposed as

$$GED(G_1, G_2) = \sum_{u \in V} GED(G_1, G_2, u) / 2, \quad (3)$$

where $GED(G_1, G_2, u)$ is the local distance between the neighborhood $N_1(u)$ of node u in G_1 and its neighborhood $N_2(u)$ in G_2 . The local distances are defined as follows:

$$GED(G_1, G_2, u) = \sum_{v \in N_1(u) \cap N_2(u)} \text{sub}(uv) + \sum_{v \in N_1(u) \setminus N_2(u)} \text{del}(uv) + \sum_{v \in N_2(u) \setminus N_1(u)} \text{ins}(uv) \quad (4)$$

Based on the local distances, we also computed cluster-level distances for a cluster of nodes $C \subseteq V$ as follows:

$$GED(G_1, G_2, C) = \sum_{u \in C} GED(G_1, G_2, u) / 2 \quad (5)$$

We used three types of edge edit cost functions, namely, uniform costs and costs based on normalized edge ranks or normalized edge weights. The uniform costs are defined by simply setting $\text{sub}(uv) = 0$ and $\text{del}(uv) = \text{ins}(uv) = 1$ for all edges uv . GED with uniform costs quantifies topological (dis-)similarity between two graphs but does not consider edge weights. Since edges are weighted in all compared diseasesomes, we additionally defined edge edit costs based on normalized weights and normalized ranks. For the normalized weights, we scaled all edge weights to the interval $[0, 1]$ via division by the maximum. For the normalized ranks, we sorted the diseasesomes' edges in increasing order with respect to their weights and then again normalized the obtained ranks to $[0, 1]$ via division by the maximum rank. Let $x_1(uv)$ be the normalized weight/rank of edge uv in diseaseome G_1 and $x_2(uv)$ be its normalized weight/rank in G_2 . Then we defined the weight-/rank-based edit costs as $\text{sub}(uv) = |x_1(uv) - x_2(uv)|$, $\text{del}(uv) = x_1(uv)$, and $\text{ins}(uv) = x_2(uv)$. That is, substitutions are expensive if the involved edge's normalized weight/rank differs a lot in the two graphs and deletions and insertions are more expensive for high-weighted/

high-ranked than for low-weighted/low-ranked edges. Since uniform, weight-based and rank-based edit costs led to similar results, we only present the results for uniform costs in the main article. Results for weight- and rank-based edit costs are shown in the supplement.

Statistical analyses based on graph edit distances

Using GED, we tested the local- and the global-scale hypotheses as follows: For each pair G_1, G_2 of compared networks, we generated 1,000 randomized counterparts G_1^1, \dots, G_1^{1000} and G_2^1, \dots, G_2^{1000} . For this, we used a random network generator which repeatedly swaps edges and non-edges to obtain randomized counterparts which exactly preserve the node degrees of the original networks^{72,73}. For each node u , we then computed $GED(G_1, G_2, u)$ as well as $GED(G_1^i, G_2^i, u)$ for each $i = 1, \dots, 1000$ and also computed the global distances $GED(G_1, G_2)$ and $GED(G_1^i, G_2^i)$.

To test the global-scale hypothesis, we computed one-sided empirical P -values as

$$P = \left(1 + \sum_{i=1}^{1000} [\text{GED}(G_1, G_2) \geq \text{GED}(G_1^i, G_2^i)] \right) / (1 + 1000), \quad (6)$$

where $[\text{true}] = 1$ and $[\text{false}] = 0$. To test the local-scale hypothesis, we used the one-sided MWU test to assess whether the local distances $\{\text{GED}(G_1, G_2, u) | u \in V\}$ for the original networks are significantly smaller than the local distances $\{\text{GED}(G_1^i, G_2^i, u) | u \in V, i = 1, \dots, 1000\}$ for the randomized counterparts. Moreover, we computed node-specific local empirical P -values as

$$P(u) = \left(1 + \sum_{i=1}^{1000} [\text{GED}(G_1, G_2, u) \geq \text{GED}(G_1^i, G_2^i, u)] \right) / (1 + 1000) \quad (7)$$

and cluster-level empirical P -values as

$$P(C) = \left(1 + \sum_{i=1}^{1000} [\text{GED}(G_1, G_2, C) \geq \text{GED}(G_1^i, G_2^i, C)] \right) / (1 + 1000) \quad (8)$$

where $C \subseteq V$ is a cluster of nodes.

Note that we consciously refrained from adjusting P -values for multiple testing. The reason for this choice is that the relevance of our results stems from the non-significance of a large fraction of the obtained P -values. If we had corrected for multiple testing, we would have inflated this fraction.

Rationale for using the graph edit distance as a measure of network dissimilarity

In addition to our version of GED, there are various other network dissimilarity measures—most notably, embedding-based^{74,75}, kernel-based⁷⁶, and message-passing-based^{77,78} approaches. We decided to use GED because, to the best of our knowledge, it is the only distance measure satisfying the following requirements necessary for our analyses:

1. To allow testing both the global- and the local-scale hypothesis, we need a graph distance measure $d(G_1, G_2)$ which is decomposable into local node distances $d(G_1, G_2, u)$.
2. The local node distances $d(G_1, G_2, u)$ should depend on u 's local neighborhoods in G_1 and G_2 but not on the overall network topologies (otherwise, we would not be testing the local-scale hypothesis when comparing local node distances).
3. Since a node alignment between the compared networks is given (disease and drug terms are aligned between the networks), both the global network distance $d(G_1, G_2)$ and the local distances $d(G_1, G_2, u)$ should be node-identity-aware rather than permutation-invariant.
4. The distances need to be computable in linear time w.r.t. the size of the networks in order to enable our large-scale permutation tests.

While most of the kernel-based methods already fall short of requirement 1, popular node-embedding-based approaches (e.g.,

node2vec⁷⁴ with subsequent distance computation in embedding space) typically do not satisfy requirements 2 through 4. Exceptions we are aware of are DeltaCon⁷⁹ (which satisfies requirements 1, 3, and 4 but not requirement 2) and the graphlet degree signature⁸⁰ (which satisfies requirements 1 and 2 but not requirements 3 and 4). Highly successful techniques in graph learning follow a message passing concept^{77,78}. When restricted to a single hop (as needed to satisfy requirement 2), these methods define node u 's embedding in the graph G_1 as $x_1(u) = g(\{l(v) | v \in N_1(u)\})$, where $l(v)$ is the label of node v (its disease or drug term) and g is a permutation-invariant function⁷⁷ mapping sets to vectors (e.g., indicator function). Here, using unique node labels renders the method node-identity-aware and allows to drop $l(\cdot)$ as a parameter of g . Such approaches fulfill all four requirements, but are essentially equivalent to GED with uniform edge costs: By comparing u 's embeddings $x_1(u)$ and $x_2(u)$, we compare the node labels of its neighboring nodes in G_1 and G_2 , which is exactly what we do with uniform GED.

Statistical analyses based on shortest path distances

We carried out analyses based on shortest path distances between (1) all disease-disease pairs in a disease-gene-gene-disease network, (2) all drug-drug pairs in a drug-protein-protein-drug network, and (3) all disease-drug pairs in a disease-protein-protein-drug network. For each network, we split the multi-set of obtained distances into multi-sets X_0 and X_1 , where X_1 contains the shortest path distances for all nodes pairs contained as edge in a reference network and X_0 contains all other shortest path distances. As reference networks, we used (1) drug-, symptom-, comorbidity-, and variant-based diseasomes, (2) a bipartite drug-indication network, and (3) an indication-based drug-drug network. We then used the one-sided MWU test to assess whether the shortest path distances contained in X_1 are significantly smaller than those contained in X_0 .

BEV network medicine is committed to the local- and the global-scale hypotheses

Recall that we have introduced BEV network medicine as the subfield of network medicine which aims at uncovering disease mechanisms by mining large-scale disease-association data. Let D_1 be data used towards this end by BEV network medicine approaches and let d_1 and d_2 be two diseases sharing an (unknown) molecular mechanisms M such that D_1 contains entries $D_1(d_1)$ and $D_1(d_2)$. If D_1 contains any useful information about disease mechanisms as assumed by BEV network medicine, M should lead to significant similarities between $D_1(d_1)$ and $D_1(d_2)$. The same holds for any other data D_2 used as input by BEV network medicine. BEV network medicine is hence implicitly committed to the claim that the edge distributions of diseasomes G_1 and G_2 constructed based on similarities in D_1 and D_2 exhibit a higher correlation than expected by chance. This, in turn, implies both the global- and the local-scale hypothesis.

Implementation

We have implemented all network analysis approaches underlying this article in a Python package called GraphSimQT. GraphSimQT uses graph-tool⁸¹ for network handling and Scipy⁸² for carrying out statistical tests and comes with all networks and scripts to reproduce the results reported in this paper. Moreover, GraphSimQT can be used to compare user-provided networks, using the techniques presented in this paper. Significance of comorbidity associations was evaluated using the Scipy implementation of Fisher's exact test and the statsmodels⁸³ implementation of Benjamini-Hochberg multiple testing correction. The GraphSimViz web tool (<https://graphsimgviz.net>) was implemented using Vue.js as a frontend framework, the Drugst.One (<https://drugst.one>) plugin as network explorer and a Django backend with a PostgreSQL database.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All networks underlying the findings of this study are available at <https://doi.org/10.5281/zenodo.7498864>. The following public databases were used to generate the networks: IID (<http://iid.ophid.utoronto.ca/>), DrugBank (<https://go.drugbank.com/>), DrugCentral (<https://drugcentral.org/>), CTD (<http://ctdbase.org/>), DisGeNET (<https://www.disgenet.org/>), OMIM (<https://omim.org/>), UniProt (<https://www.uniprot.org/>), MONDO (<https://MONDO.monarchinitiative.org/>), NeDRex (<https://nedrex.net/>), and HPO (<https://hpo.jax.org/app/>). Version numbers of all used databases can be found in an AIME report⁸⁴ for our study (<https://aime.report/6bdnlg>). The comorbidity-based diseasome was constructed based on data provided by the Estonian Biobank (<https://genomics.ut.ee/en/content/estonian-biobank>, available from the Estonian Biobank upon request). The construction of the networks is described in the Methods section of this paper. Our study is based on public databases (including DisGeNET, OMIM, DrugBank, HPO, and more) which do not contain sex-specific information. Therefore, no sex-specific analyses could be carried out. Source data are provided in this paper. They can also be downloaded from https://api.graphsimviz.net/download_results. Source data are provided with this paper.

Code availability

The GraphSimQT tool is available at <https://github.com/repotrial/graphsimqt>, together with scripts to reproduce all results reported in this article. A stable version is available from Zenodo⁸⁵ (<https://doi.org/10.5281/zenodo.7498864>). The source code of the frontend and the backend of GraphSimViz is available at <https://github.com/repotrial/GraphSimViz-frontend> and <https://github.com/repotrial/GraphSimViz-backend>, respectively.

References

- Goh, K.-I. et al. The human disease network. *Proc. Natl. Acad. Sci. USA*. **104**, 8685–8690 (2007).
- Barabási, A.-L., Gulbahce, N. & Loscalzo, J. Network medicine: A network-based approach to human disease. *Nat. Rev. Genet.* **12**, 56–68 (2011).
- Baumbach, J. & Schmidt, H. H. W. The end of medicine as we know it: Introduction to the new journal, systems medicine. *Syst. Med.* **1**, 1–2 (2018).
- Maron, B. A. et al. A global network for network medicine. *NPJ Syst. Biol. Appl.* **6**, 29 (2020).
- Nogales, C. et al. Network pharmacology: Curing causal mechanisms instead of treating symptoms. *Trends Pharmacol. Sci.* **43**, 136–150 (2022).
- Loscalzo, J., Kohane, I. & Barabasi, A.-L. Human disease classification in the postgenomic era: A complex systems approach to human pathobiology. *Mol. Syst. Biol.* **3**, 124 (2007).
- Agache, I. & Akdis, C. A. Precision medicine and phenotypes, endotypes, genotypes, regiotypes, and theratypes of allergic diseases. *J. Clin. Invest.* **129**, 1493–1503 (2019).
- Anderson, G. P. Endotyping asthma: New insights into key pathogenic mechanisms in a complex, heterogeneous disease. *Lancet* **372**, 1107–1119 (2008).
- Lötvall, J. et al. Asthma endotypes: A new approach to classification of disease entities within the asthma syndrome. *J. Allergy Clin. Immunol.* **127**, 355–360 (2011).
- Ghiassian, S. D. et al. Endophenotype network models: Common core of complex diseases. *Sci. Rep.* **6**, 27414 (2016).

11. Leopold, J. A., Maron, B. A. & Loscalzo, J. The application of big data to cardiovascular disease: Paths to precision medicine. *J. Clin. Invest.* **130**, 29–38 (2020).
12. Sharma, A. et al. Controllability in an islet-specific regulatory network identifies the transcriptional factor NFATC4, which regulates Type 2 Diabetes-associated genes. *NPJ Syst. Biol. Appl.* **4**, 25 (2018).
13. AbdulHameed, M. D. M. et al. Systems level analysis and identification of pathways and networks associated with liver fibrosis. *PLoS One* **9**, e112193 (2014).
14. Samokhin, A. O. et al. NEDD9 targets COL3A1 to promote endothelial fibrosis and pulmonary arterial hypertension. *Sci. Transl. Med.* **10**, eaap7294 (2018).
15. Sharma, A. et al. A disease module in the interactome explains disease heterogeneity, drug response and captures novel pathways and genes in asthma. *Hum. Mol. Genet.* **24**, 3005–3020 (2015).
16. Maron, B. A. et al. Individualized interactomes for network-based precision medicine in hypertrophic cardiomyopathy with implications for other clinical pathophenotypes. *Nat. Commun.* **12**, 873 (2021).
17. Mirzakhani, H. et al. Early pregnancy vitamin D status and risk of preeclampsia. *J. Clin. Invest.* **126**, 4702–4715 (2016).
18. Halu, A. et al. Exploring the cross-phenotype network region of disease modules reveals concordant and discordant pathways between chronic obstructive pulmonary disease and idiopathic pulmonary fibrosis. *Hum. Mol. Genet.* **28**, 2352–2364 (2019).
19. Menche, J. et al. Disease networks. Uncovering disease-disease relationships through the incomplete interactome. *Science* **347**, 1257601 (2015).
20. Iida, M., Iwata, M. & Yamanishi, Y. Network-based characterization of disease-disease relationships in terms of drugs and therapeutic targets. *Bioinformatics* **36**, i516–i524 (2020).
21. Guney, E., Menche, J., Vidal, M. & Barabási, A.-L. Network-based in silico drug efficacy screening. *Nat. Commun.* **7**, 10331 (2016).
22. Cheng, F. et al. Network-based approach to prediction and population-based validation of in silico drug repurposing. *Nat. Commun.* **9**, 2691 (2018).
23. Cheng, F., Kovács, I. A. & Barabási, A.-L. Network-based prediction of drug combinations. *Nat. Commun.* **10**, 1197 (2019).
24. Zhou, Y. et al. Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. *Cell Disco.* **6**, 14 (2020).
25. Schaefer, M. H., Serrano, L. & Andrade-Navarro, M. A. Correcting for the study bias associated with protein-protein interaction measurements reveals differences between protein degree distributions from different cancer types. *Front. Genet.* **6**, 260 (2015).
26. Wachi, S., Yoneda, K. & Wu, R. Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. *Bioinformatics* **21**, 4205–4208 (2005).
27. Jonsson, P. F. & Bates, P. A. Global topological features of cancer proteins in the human interactome. *Bioinformatics* **22**, 2291–2297 (2006).
28. Rambaldi, D., Giorgi, F. M., Capuani, F., Ciliberto, A. & Ciccarelli, F. D. Low duplicability and network fragility of cancer genes. *Trends Genet.* **24**, 427–430 (2008).
29. Lazareva, O., Baumbach, J., List, M. & Blumenthal, D. B. On the limits of active module identification. *Brief. Bioinform.* **22**, bbab066 (2021).
30. Haynes, W. A., Tomczak, A. & Khatri, P. Gene annotation bias impedes biomedical research. *Sci. Rep.* **8**, 1362 (2018).
31. Gene Ontology Consortium. The gene ontology resource: Enriching a GOld mine. *Nucleic Acids Res.* **49**, D325–D334 (2021).
32. Kustatscher, G. et al. Understudied proteins: Opportunities and challenges for functional proteomics. *Nat. Methods* **19**, 774–779 (2022).
33. Stoeger, T., Gerlach, M., Morimoto, R. I. & Nunes Amaral, L. A. Large-scale investigation of the reasons why potentially important genes are ignored. *PLoS Biol.* **16**, e2006643 (2018).
34. Rodriguez-Esteban, R. The speed of information propagation in the scientific network distorts biomedical research. *PeerJ.* **10**, e12764 (2022).
35. Langhauser, F. et al. A disease cluster-based drug repurposing of soluble guanylate cyclase activators from smooth muscle relaxation to direct neuroprotection. *npj Syst. Biol. Appl.* **4**, 1–13 (2018).
36. Piñero, J. et al. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.* **48**, D845–D855 (2020).
37. Avram, S. et al. DrugCentral 2021 supports drug discovery and repositioning. *Nucleic Acids Res.* **49**, D1160–D1169 (2021).
38. Sanfeliu, A. & Fu, K.-S. A distance measure between attributed relational graphs for pattern recognition. *IEEE Trans. Syst. Man Cybern.* **13**, 353–362 (1983).
39. Bunke, H. & Allermann, G. Inexact graph matching for structural pattern recognition. *Pattern Recognit. Lett.* **1**, 245–253 (1983).
40. Vasilevsky, N. A. et al. Mondo: Unifying diseases for the world, by the world. *medRxiv.* 2022.04.13.22273750 <https://doi.org/10.1101/2022.04.13.22273750> (2022).
41. Bodenreider, O. The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Res.* **32**, D267–D270 (2004).
42. World Health Organization. *The International Statistical Classification of Diseases and Health Related Problems ICD-10: Tenth Revision. Volume 2: Instruction Manual.* (World Health Organization, 2004).
43. Amberger, J. S., Bocchini, C. A., Scott, A. F. & Hamosh, A. OMIM.org: leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Res.* **47**, D1038–D1043 (2019).
44. Nogales, C. et al. Network pharmacology: curing causal mechanisms instead of treating symptoms. *Trends Pharmacol. Sci.* <https://doi.org/10.1016/j.tips.2021.11.004> (2021).
45. Aguirre-Plans, J. et al. GUILDify v2.0: A tool to identify molecular networks underlying human diseases, their comorbidities and their druggable targets. *J. Mol. Biol.* **431**, 2477–2484 (2019).
46. Himmelstein, D. S. et al. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *Elife* **6**, e26726 (2017).
47. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).
48. Davis, A. P. et al. Comparative Toxicogenomics Database (CTD): Update 2021. *Nucleic Acids Res.* **49**, D1138–D1143 (2021).
49. Hivert, V., Martin, N., Hanauer, M. & Aymé, S. New functionalities in Orphanet for orphan drugs, R&D and marketing authorisations to better serve the rare diseases community. *Orphanet J. Rare Dis.* **5**, <https://doi.org/10.1186/1750-1172-5-s1-p25> (2010).
50. Rehm, H. L. et al. ClinGen — The clinical genome resource. *N. Engl. J. Med.* **372**, 2235–2242 (2015).
51. Martin, A. R. et al. PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nat. Genet.* **51**, 1560–1565 (2019).
52. Tamborero, D. et al. Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Med.* **10**, 25 (2018).
53. Gutiérrez-Sacristán, A. et al. PsyGeNET: A knowledge platform on psychiatric disorders and their genes. *Bioinformatics* **31**, 3075–3077 (2015).
54. Li, M. J. et al. GWASdb v2: An update database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res.* **44**, D869–D876 (2016).
55. Landrum, M. J. & Kattman, B. L. ClinVar at five years: Delivering on the promise. *Hum. Mutat.* **39**, 1623–1630 (2018).

56. MacArthur, J. et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2017).
57. Bravo, À., Piñero, J., Queralt, N., Rautschka, M. & Furlong, L. I. Extraction of relations between genes and diseases from text and large-scale data analysis: Implications for translational research. *BMC Bioinformatics.* **16**, 55 (2015).
58. Capell, B. C. & Collins, F. S. Human laminopathies: Nuclei gone genetically awry. *Nat. Rev. Genet.* **7**, 940–952 (2006).
59. Cantor, R. M., Lange, K. & Sinsheimer, J. S. Prioritizing GWAS results: A review of statistical methods and recommendations for their application. *Am. J. Hum. Genet.* **86**, 6–22 (2010).
60. Larsen, S. J., Schmidt, H. H. W. & Baumbach, J. De Novo and supervised endophenotyping using network-guided ensemble learning. *Syst. Med.* **3**, 8–21 (2020).
61. Lazareva, O. et al. BiCoN: Network-constrained biclustering of patients and omics data. *Bioinformatics* **37**, 2398–2404 (2020).
62. Zolotareva, O. et al. Identification of differentially expressed gene modules in heterogeneous diseases. *Bioinformatics* <https://doi.org/10.1093/bioinformatics/btaa1038> (2020).
63. List, M. et al. KeyPathwayMinerWeb: Online multi-omics network enrichment. *Nucleic Acids Res.* **44**, W98–W104 (2016).
64. Batra, R. et al. On the performance of de novo pathway enrichment. *NPJ Syst. Biol. Appl.* **3**, 6 (2017).
65. Ghiassian, S. D., Menche, J. & Barabási, A.-L. A Disease Module Detection (DIAMOND) Algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. *PLoS Comput. Biol.* **11**, e1004120 (2015).
66. Levi, H., Elkon, R. & Shamir, R. DOMINO: A network-based active module identification algorithm with reduced rate of false calls. *Mol. Syst. Biol.* **17**, e9593 (2021).
67. Bennett, J. et al. Robust disease module mining via enumeration of diverse prize-collecting Steiner trees. *Bioinformatics* **38**, 1600–1606 (2022).
68. Sadegh, S. et al. Network medicine for disease module identification and drug repurposing with the NeDRex platform. *Nat. Commun.* **12**, 6848 (2021).
69. National Library of Medicine (U.S.). *Medical Subject Headings: Main Headings, Subheadings and Cross References Used in the Index Medicus and the National Library of Medicine Catalog.* (1960).
70. Blumenthal, D. B., Boria, N., Gamper, J., Bougleux, S. & Brun, L. Comparing heuristics for graph edit distance computation. *Vldb J.* **29**, 419–458 (2020).
71. Blumenthal, D. B. & Gamper, J. On the exact computation of the graph edit distance. *Pattern Recognit. Lett.* **134**, 46–57 (2020).
72. Gkantsidis, C., Mihail, M. & Zegura, E. W. The Markov chain simulation method for generating connected power law random graphs. in *ALLENEX 2003* (ed. Ladner, R. E.) 16–25 (SIAM, 2003).
73. Viger, F. & Latapy, M. Efficient and simple generation of random simple connected graphs with prescribed degree sequence. *J. Complex Netw.* **4**, 15–37 (2016).
74. Grover, A. & Leskovec, J. node2vec: Scalable Feature Learning for Networks. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016* (eds. Krishnapuram, B. et al.) 855–864 (ACM, 2016).
75. Rossi, R. A. et al. On proximity and structural role-based embeddings in networks: Misconceptions, techniques, and applications. *ACM Trans. Knowl. Discov. Data* **14**, 1–37 (2020).
76. Borgwardt, K., Ghisu, E., Llinares-López, F., O’Bray, L. & Rieck, B. Graph Kernels: State-of-the-art and future challenges. *Found. Trends® Mach. Learn.* **13**, 531–712 (2020).
77. Morris, C. et al. Weisfeiler and Leman go Machine Learning: The Story so far. *arXiv [cs.LG] Preprint* at <https://doi.org/10.48550/arXiv.2112.09992> (2021).
78. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. Neural Message Passing for Quantum Chemistry. in *Proceedings of the 34th International Conference on Machine Learning* (eds. Precup, D. & Teh, Y. W.) vol. 70 1263–1272 (PMLR, 06-11 Aug 2017).
79. Koutra, D., Shah, N., Vogelstein, J. T., Gallagher, B. & Faloutsos, C. DeltaCon: Principled massive-graph similarity function with attribution. *ACM Trans. Knowl. Discov. Data* **10**, 1–43 (2016).
80. Przulj, N. Biological network comparison using graphlet degree distribution. *Bioinformatics* **23**, e177–e183 (2007).
81. Peixoto, T. P. The graph-tool python library. *figshare* <https://doi.org/10.6084/m9.figshare.1164194> (2014).
82. Virtanen, P. et al. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
83. Seabold, S. & Perktold, J. Statsmodels: Econometric and Statistical Modeling with Python. In *Proceedings of the 9th Python in Science Conference.* 92–96. <https://doi.org/10.25080/Majora-92bf1922-011> (2010).
84. Matschinske, J. et al. The AIMe registry for artificial intelligence in biomedical research. *Nat. Methods* **18**, 1128–1131 (2021).
85. Sadegh, S. et al. Lacking mechanistic disease definitions and corresponding association data hamper progress in network medicine and beyond, *reprotrial/graphsimqt: GraphSimQT.* <https://doi.org/10.5281/zenodo.7498864> (2023).
86. Köhler, S. et al. The human phenotype ontology in 2021. *Nucleic Acids Res.* **49**, D1207–D1217 (2021).
87. Lee, D., de Keizer, N., Lau, F. & Cornet, R. Literature review of SNOMED CT use. *J. Am. Med. Inform. Assoc.* **21**, e11–e19 (2014).
88. Wishart, D. S. et al. DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Res.* **46**, D1074–D1082 (2018).
89. Kotlyar, M., Pastrello, C., Malik, Z. & Jurisica, I. IID 2018 update: Context-specific physical protein-protein interactions in human, model organisms and domesticated species. *Nucleic Acids Res.* **47**, D581–D589 (2019).
90. Leitsalu, L. et al. Cohort profile: Estonian biobank of the estonian genome center, University of Tartu. *Int. J. Epidemiol.* **44**, 1137–1147 (2015).

Acknowledgements

S.S., E.A., J.S., A.W., and J.B. are grateful for financial support from REPO-TRIAL. REPO-TRIAL has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 777111. This publication reflects only the authors’ view and the European Commission is not responsible for any use that may be made of the information it contains. This work was supported by the German Federal Ministry of Education and Research (BMBF) within the framework of the e:Med research and funding concept (grant O1ZX1908A) (S.S. and J.B.). J.B. was partially funded by his VILLUM Young Investigator Grant no. 13154. N.M.K. was supported by the Vienna Science and Technology Fund (WWTF) through project VRG19-009. The work of J.K. was supported by the Estonian Research Council grant PUT (PRG1291). T.H. was supported by the European Union through the European Regional Development Fund (Project No. 2014-2020.4.01.15-0012). The Estonian Biobank was funded by the European Union through the European Regional Development Fund Project No. 2014-2020.4.01.15-0012 GEN-TRANSMED. Data analysis by J.K. and T.H. was carried out in part in the High-Performance Computing Center of University of Tartu. The Estonian Biobank Research Team includes Mari Nelis, Lili Milani, Tõnu Esko, Andres Metspalu, and Reedik Mägi. Figures 1–6 and Supplementary Fig. 4 and 12 were created with BioRender.com.

Author contributions

D.B.B. and S.S. conceived and designed this study and implemented the GraphSimQT Python package to compare the different networks. S.S. carried out the analyses. J.S., S.S., and K.A. integrated the data and constructed the networks. A.Ma. implemented the GraphSimViz web

tool. D.B.B., S.S., E.A., N.M.K., and A.Mo. drafted the manuscript. J.K., T.H., and the Estonian Biobank Research Team provided the comorbidity data. D.B.B., J.B., and A.W. supervised the project. All authors provided critical feedback and discussion and assisted in interpreting the results and writing the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at

<https://doi.org/10.1038/s41467-023-37349-4>.

Correspondence and requests for materials should be addressed to David B. Blumenthal.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

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

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

© The Author(s) 2023

Acronyms

DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
COVID-19	coronavirus disease 2019
CoVex	CoronaVirus Explorer
MuST	Multi-Steiner Tree
mRNA	Messenger ribonucleic acid
3D	three-dimensional
miRNA	Micro ribonucleic acid
OMIM	Online Mendelian Inheritance in Man
SNV	Single-Nucleotide Variation
SNP	Single-Nucleotide Polymorphism
SV	Structural Variation
CNV	Copy Number Variation
GWAS	Genome-Wide Association Studies
NGS	Next-Generation Sequencing
WGS	Whole-Genome Sequencing
AS	Alternative Splicing
GEO	Gene Expression Omnibus
MS	Mass Spectrometry
ELISA	Enzyme-Linked Immunosorbent Assay
PPI	Protein-Protein Interaction
TAP	Tandem Affinity Purification
Y2H	Yeast Two-Hybrid
GO	Gene Ontology
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2
AP	Affinity-Purification
R&D	Research and Development
HTS	High-Throughput Screening
KEGG	Kyoto Encyclopedia of Genes and Genomes
DAG	Directed Acyclic Graph

MONDO Mondo Disease Ontology
GDA Gene-Disease Association
EHR Electronic Health Record
IID Integrated Interactions Database
GSORA Gene Set Over-Representation Analysis
GSEA Gene Set Enrichment Analysis
DMI Disease Module Identification
DNE De novo Network Enrichment
ASM Aggregate Score Methods
ST Steiner Trees
SPM Score Propagation Methods
MC Module Cover
BiCoN Biclustering Constrained by Networks
ICD International Classification of Diseases
MeSH Medical Subject Headings
UMLS CUI Unified Medical Language System Concept Unique Identifier
SNOMED CT Systematized Nomenclature of Medicine Clinical Term
BEV Bird's-Eye-View
HGNC HUGO Gene Nomenclature Committee
ORDO Orphanet Rare Disease Ontology
REST Representational State Transfer
FTP File Transfer Protocol
SQL Structured Query Language
OSGi Open Service Gateway Initiative
ST Steiner Tree
DCG Discounted Cumulative Gain
GED Graph Edit Distance
MWU Mann-Whitney U
GraphSimQT Graph Similarity Quantification Tool
GraphSimViz Graph Similarity Visualizer

References

1. Ginsburg GS, Phillips KA. Precision Medicine: From Science To Value. *Health Aff* . 2018;37: 694–701. doi:10.1377/hlthaff.2017.1624
2. Garrido-Rodríguez M, Zirngibl K, Ivanova O, Lobentanzer S, Saez-Rodríguez J. Integrating knowledge and omics to decipher mechanisms via large-scale models of signaling networks. *Mol Syst Biol*. 2022;18: e11036. doi:10.15252/msb.202211036
3. Hood L, Flores M. A personal view on systems medicine and the emergence of proactive P4 medicine: predictive, preventive, personalized and participatory. *N Biotechnol*. 2012;29: 613–624. doi:10.1016/j.nbt.2012.03.004
4. Sharma A, Menche J, Huang CC, Ort T, Zhou X, Kitsak M, et al. A disease module in the interactome explains disease heterogeneity, drug response and captures novel pathways and genes in asthma. *Hum Mol Genet*. 2015;24: 3005–3020. doi:10.1093/hmg/ddv001
5. Zanzoni A, Soler-López M, Aloy P. A network medicine approach to human disease. *FEBS Lett*. 2009;583: 1759–1765. doi:10.1016/j.febslet.2009.03.001
6. Ashburn TT, Thor KB. Drug repositioning: identifying and developing new uses for existing drugs. *Nat Rev Drug Discov*. 2004;3: 673–683. doi:10.1038/nrd1468
7. Krishnamurthy N, Grimshaw AA, Axson SA, Choe SH, Miller JE. Drug repurposing: a systematic review on root causes, barriers and facilitators. *BMC Health Serv Res*. 2022;22: 970. doi:10.1186/s12913-022-08272-z
8. Rodrigues L, Bento Cunha R, Vassilevskaia T, Viveiros M, Cunha C. Drug Repurposing for COVID-19: A Review and a Novel Strategy to Identify New Targets and Potential Drug Candidates. *Molecules*. 2022;27. doi:10.3390/molecules27092723
9. Selvaraj N, Swaroop AK, Nidamanuri BSS, Kumar RR, Natarajan J, Selvaraj J. Network-based Drug Repurposing: A Critical Review. *Curr Drug Res Rev*. 2022;14: 116–131. doi:10.2174/2589977514666220214120403
10. Fiscon G, Conte F, Farina L, Paci P. A Comparison of Network-Based Methods for Drug Repurposing along with an Application to Human Complex Diseases. *Int J Mol Sci*. 2022;23. doi:10.3390/ijms23073703
11. Sadegh S, Matschinske J, Blumenthal DB, Galindez G, Kacprowski T, List M, et al. Exploring the SARS-CoV-2 virus-host-drug interactome for drug repurposing. *Nat Commun*. 2020;11: 3518. doi:10.1038/s41467-020-17189-2
12. Sadegh S, Skelton J, Anastasi E, Bernett J, Blumenthal DB, Galindez G, et al. Network medicine for disease module identification and drug repurposing with the NeDRex platform. *Nat Commun*. 2021;12: 6848. doi:10.1038/s41467-

13. Galindez G, Matschinske J, Rose TD, Sadegh S, Salgado-Albarrán M, Späth J, et al. Lessons from the COVID-19 pandemic for advancing computational drug repurposing strategies. *Nat Comput Sci.* 2021;1: 33–41. doi:10.1038/s43588-020-00007-6
14. Matschinske J, Salgado-Albarrán M, Sadegh S, Bongiovanni D, Baumbach J, Blumenthal DB. Individuating Possibly Repurposable Drugs and Drug Targets for COVID-19 Treatment Through Hypothesis-Driven Systems Medicine Using CoVex. *Assay Drug Dev Technol.* 2020;18: 348–355. doi:10.1089/adt.2020.1010
15. Bernett J, Krupke D, Sadegh S, Baumbach J, Fekete SP, Kacprowski T, et al. Robust disease module mining via enumeration of diverse prize-collecting Steiner trees. *Bioinformatics.* 2022;38: 1600–1606. doi:10.1093/bioinformatics/btab876
16. Galindez G, Sadegh S, Baumbach J, Kacprowski T, List M. Network-based approaches for modeling disease regulation and progression. *Comput Struct Biotechnol J.* 2023;21: 780–795. doi:10.1016/j.csbj.2022.12.022
17. Cobb M. 60 years ago, Francis Crick changed the logic of biology. *PLoS Biol.* 2017;15: e2003243. doi:10.1371/journal.pbio.2003243
18. Crick F. Central dogma of molecular biology. *Nature.* 1970;227: 561–563. doi:10.1038/227561a0
19. Egger G, Liang G, Aparicio A, Jones PA. Epigenetics in human disease and prospects for epigenetic therapy. *Nature.* 2004;429: 457–463. doi:10.1038/nature02625
20. Robinson VL. Rethinking the central dogma: noncoding RNAs are biologically relevant. *Urol Oncol.* 2009;27: 304–306. doi:10.1016/j.urolonc.2008.11.004
21. Katsonis P, Koire A, Wilson SJ, Hsu T-K, Lua RC, Wilkins AD, et al. Single nucleotide variations: biological impact and theoretical interpretation. *Protein Sci.* 2014;23: 1650–1666. doi:10.1002/pro.2552
22. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* 2005;33: D514–7. doi:10.1093/nar/gki033
23. Badano JL, Katsanis N. Beyond Mendel: an evolving view of human genetic disease transmission. *Nat Rev Genet.* 2002;3: 779–789. doi:10.1038/nrg910
24. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature.* 2004;431: 931–945. doi:10.1038/nature03001
25. Gonzaga-Jauregui C, Lupski JR, Gibbs RA. Human genome sequencing in health and disease. *Annu Rev Med.* 2012;63: 35–61. doi:10.1146/annurev-

med-051010-162644

26. Sukhumsirichart W. Polymorphisms. Genetic Diversity and Disease Susceptibility. InTech; 2018. doi:10.5772/intechopen.76728
27. Suwinski P, Ong C, Ling MHT, Poh YM, Khan AM, Ong HS. Advancing Personalized Medicine Through the Application of Whole Exome Sequencing and Big Data Analytics. *Front Genet.* 2019;10: 49. doi:10.3389/fgene.2019.00049
28. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, et al. Global variation in copy number in the human genome. *Nature.* 2006;444: 444–454. doi:10.1038/nature05329
29. Zarrei M, MacDonald JR, Merico D, Scherer SW. A copy number variation map of the human genome. *Nat Rev Genet.* 2015;16: 172–183. doi:10.1038/nrg3871
30. Gross AM, Ajay SS, Rajan V, Brown C, Bluske K, Burns NJ, et al. Copy-number variants in clinical genome sequencing: deployment and interpretation for rare and undiagnosed disease. *Genet Med.* 2019;21: 1121–1130. doi:10.1038/s41436-018-0295-y
31. Ponomarenko EA, Poverennaya EV, Ilgisonis EV, Pyatnitskiy MA, Kopylov AT, Zgoda VG, et al. The Size of the Human Proteome: The Width and Depth. *Int J Anal Chem.* 2016;2016: 7436849. doi:10.1155/2016/7436849
32. Maier T, Güell M, Serrano L. Correlation of mRNA and protein in complex biological samples. *FEBS Lett.* 2009;583: 3966–3973. doi:10.1016/j.febslet.2009.10.036
33. Rogers S, Girolami M, Kolch W, Waters KM, Liu T, Thrall B, et al. Investigating the correspondence between transcriptomic and proteomic expression profiles using coupled cluster models. *Bioinformatics.* 2008;24: 2894–2900. doi:10.1093/bioinformatics/btn553
34. Ghazalpour A, Bennett B, Petyuk VA, Orozco L, Hagopian R, Mungrue IN, et al. Comparative analysis of proteome and transcriptome variation in mouse. *PLoS Genet.* 2011;7: e1001393. doi:10.1371/journal.pgen.1001393
35. Wang Z, Lachmann A, Ma'ayan A. Mining data and metadata from the gene expression omnibus. *Biophys Rev.* 2019;11: 103–110. doi:10.1007/s12551-018-0490-8
36. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res.* 2013;41: D991–5. doi:10.1093/nar/gks1193
37. Gao GF, Parker JS, Reynolds SM, Silva TC, Wang L-B, Zhou W, et al. Before and After: Comparison of Legacy and Harmonized TCGA Genomic Data Commons' Data. *Cell Syst.* 2019;9: 24–34.e10. doi:10.1016/j.cels.2019.06.006
38. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 2014;15: 550.

doi:10.1186/s13059-014-0550-8

39. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 2015;43: e47. doi:10.1093/nar/gkv007
40. Stuart JM, Segal E, Koller D, Kim SK. A gene-coexpression network for global discovery of conserved genetic modules. *Science.* 2003;302: 249–255. doi:10.1126/science.1087447
41. Wilkerson MD, Yin X, Hoadley KA, Liu Y, Hayward MC, Cabanski CR, et al. Lung squamous cell carcinoma mRNA expression subtypes are reproducible, clinically important, and correspond to normal cell types. *Clin Cancer Res.* 2010;16: 4864–4875. doi:10.1158/1078-0432.CCR-10-0199
42. Bailey P, Chang DK, Nones K, Johns AL, Patch A-M, Gingras M-C, et al. Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature.* 2016;531: 47–52. doi:10.1038/nature16965
43. Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol.* 2009;27: 1160–1167. doi:10.1200/JCO.2008.18.1370
44. Lazareva O, Canzar S, Yuan K, Baumbach J, Blumenthal DB, Tieri P, et al. BiCoN: Network-constrained biclustering of patients and omics data. *Bioinformatics.* 2020. doi:10.1093/bioinformatics/btaa1076
45. Keren H, Lev-Maor G, Ast G. Alternative splicing and evolution: diversification, exon definition and function. *Nat Rev Genet.* 2010;11: 345–355. doi:10.1038/nrg2776
46. Halperin RF, Hegde A, Lang JD, Raupach EA, C4RCD Research Group, Legendre C, et al. Improved methods for RNAseq-based alternative splicing analysis. *Sci Rep.* 2021;11: 10740. doi:10.1038/s41598-021-89938-2
47. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet.* 2008;40: 1413–1415. doi:10.1038/ng.259
48. Tan H, Bao J, Zhou X. Genome-wide mutational spectra analysis reveals significant cancer-specific heterogeneity. *Sci Rep.* 2015;5: 12566. doi:10.1038/srep12566
49. Kochetov AV. Alternative translation start sites and hidden coding potential of eukaryotic mRNAs. *Bioessays.* 2008;30: 683–691. doi:10.1002/bies.20771
50. Rolland T, Taşan M, Charlotteaux B, Pevzner SJ, Zhong Q, Sahni N, et al. A proteome-scale map of the human interactome network. *Cell.* 2014;159: 1212–1226. doi:10.1016/j.cell.2014.10.050
51. Link AJ, Eng J, Schieltz DM, Carmack E, Mize GJ, Morris DR, et al. Direct analysis of protein complexes using mass spectrometry. *Nat Biotechnol.* 1999;17: 676–682. doi:10.1038/10890

52. Taylor SW, Fahy E, Ghosh SS. Global organellar proteomics. *Trends Biotechnol.* 2003;21: 82–88. doi:10.1016/S0167-7799(02)00037-9
53. Goldberg AL. Protein degradation and protection against misfolded or damaged proteins. *Nature.* 2003;426: 895–899. doi:10.1038/nature02263
54. Hochstrasser M. Ubiquitin-dependent protein degradation. *Annu Rev Genet.* 1996;30: 405–439. doi:10.1146/annurev.genet.30.1.405
55. Harper JW, Bennett EJ. Proteome complexity and the forces that drive proteome imbalance. *Nature.* 2016;537: 328–338. doi:10.1038/nature19947
56. Wu CH, Chen C, editors. *Bioinformatics for Comparative Proteomics.* 2011th ed. New York, NY: Humana Press; 2010. doi:10.1007/978-1-60761-977-2
57. Ali A, Bagchi A. An overview of protein-protein interaction. *Curr Chem Biol.* 2015;9: 53–65. doi:10.2174/221279680901151109161126
58. Nooren IMA, Thornton JM. Diversity of protein-protein interactions. *EMBO J.* 2003;22: 3486–3492. doi:10.1093/emboj/cdg359
59. Bagchi A. Prediction of protein-protein interactions. *Computational Intelligence and Pattern Analysis in Biological Informatics.* Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2010. pp. 325–347. doi:10.1002/9780470872352.ch15
60. Ivanic J, Yu X, Wallqvist A, Reifman J. Influence of protein abundance on high-throughput protein-protein interaction detection. *PLoS One.* 2009;4: e5815. doi:10.1371/journal.pone.0005815
61. Rao VS, Srinivas K, Sujini GN, Kumar GNS. Protein-protein interaction detection: methods and analysis. *Int J Proteomics.* 2014;2014: 147648. doi:10.1155/2014/147648
62. Kotlyar M, Pastrello C, Pivetta F, Lo Sardo A, Cumbaa C, Li H, et al. In silico prediction of physical protein interactions and characterization of interactome orphans. *Nat Methods.* 2015;12: 79–84. doi:10.1038/nmeth.3178
63. Zhang QC, Petrey D, Deng L, Qiang L, Shi Y, Thu CA, et al. Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature.* 2012;490: 556–560. doi:10.1038/nature11503
64. Elefsinioti A, Saraç ÖS, Hegele A, Plake C, Hubner NC, Poser I, et al. Large-scale de novo prediction of physical protein-protein association. *Mol Cell Proteomics.* 2011;10: M111.010629. doi:10.1074/mcp.M111.010629
65. Gordon DE, Jang GM, Bouhaddou M, Xu J, Obernier K, White KM, et al. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature.* 2020;583: 459–468. doi:10.1038/s41586-020-2286-9
66. Scannell JW, Blanckley A, Boldon H, Warrington B. Diagnosing the decline in pharmaceutical R&D efficiency. *Nat Rev Drug Discov.* 2012;11: 191–200. doi:10.1038/nrd3681

67. Emilien G, Ponchon M, Caldas C, Isacson O, Maloteaux JM. Impact of genomics on drug discovery and clinical medicine. *QJM*. 2000;93: 391–423. doi:10.1093/qjmed/93.7.391
68. Joachimiak A. High-throughput crystallography for structural genomics. *Curr Opin Struct Biol*. 2009;19: 573–584. doi:10.1016/j.sbi.2009.08.002
69. Mayr LM, Fuerst P. The future of high-throughput screening. *J Biomol Screen*. 2008;13: 443–448. doi:10.1177/1087057108319644
70. Ringel MS, Scannell JW, Baedeker M, Schulze U. Breaking Eroom's Law. *Nat Rev Drug Discov*. 2020;19: 833–834. doi:10.1038/d41573-020-00059-3
71. Pushpakom S, Iorio F, Eyers PA, Escott KJ, Hopper S, Wells A, et al. Drug repurposing: progress, challenges and recommendations. *Nat Rev Drug Discov*. 2019;18: 41–58. doi:10.1038/nrd.2018.168
72. Park K. A review of computational drug repurposing. *Transl Clin Pharmacol*. 2019;27: 59–63. doi:10.12793/tcp.2019.27.2.59
73. Breckenridge A, Jacob R. Overcoming the legal and regulatory barriers to drug repurposing. *Nat Rev Drug Discov*. 2019;18: 1–2. doi:10.1038/nrd.2018.92
74. Nosengo N. Can you teach old drugs new tricks? *Nature*. 2016;534: 314–316. doi:10.1038/534314a
75. Patrono C. Aspirin as an antiplatelet drug. *N Engl J Med*. 1994;330: 1287–1294. doi:10.1056/NEJM199405053301808
76. Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, Hufeisen SJ, et al. Predicting new molecular targets for known drugs. *Nature*. 2009;462: 175–181. doi:10.1038/nature08506
77. Hieronymus H, Lamb J, Ross KN, Peng XP, Clement C, Rodina A, et al. Gene expression signature-based chemical genomic prediction identifies a novel class of HSP90 pathway modulators. *Cancer Cell*. 2006;10: 321–330. doi:10.1016/j.ccr.2006.09.005
78. Dudley JT, Deshpande T, Butte AJ. Exploiting drug-disease relationships for computational drug repositioning. *Brief Bioinform*. 2011;12: 303–311. doi:10.1093/bib/bbr013
79. Iorio F, Rittman T, Ge H, Menden M, Saez-Rodriguez J. Transcriptional data: a new gateway to drug repositioning? *Drug Discov Today*. 2013;18: 350–357. doi:10.1016/j.drudis.2012.07.014
80. Chiang AP, Butte AJ. Systematic evaluation of drug-disease relationships to identify leads for novel drug uses. *Clin Pharmacol Ther*. 2009;86: 507–510. doi:10.1038/clpt.2009.103
81. Iorio F, Bosotti R, Scacheri E, Belcastro V, Mithbaokar P, Ferriero R, et al. Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proc Natl Acad Sci U S A*. 2010;107: 14621–14626.

doi:10.1073/pnas.1000138107

82. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*. 2006;313: 1929–1935. doi:10.1126/science.1132939
83. Oprea TI, Tropsha A, Faulon J-L, Rintoul MD. Systems chemical biology. *Nat Chem Biol*. 2007;3: 447–450. doi:10.1038/nchembio0807-447
84. Li YY, An J, Jones SJM. A computational approach to finding novel targets for existing drugs. *PLoS Comput Biol*. 2011;7: e1002139. doi:10.1371/journal.pcbi.1002139
85. Sanseau P, Agarwal P, Barnes MR, Pastinen T, Richards JB, Cardon LR, et al. Use of genome-wide association studies for drug repositioning. *Nat Biotechnol*. 2012;30: 317–320. doi:10.1038/nbt.2151
86. Greene CS, Voight BF. Pathway and network-based strategies to translate genetic discoveries into effective therapies. *Hum Mol Genet*. 2016;25: R94–R98. doi:10.1093/hmg/ddw160
87. Guney E, Menche J, Vidal M, Barabási A-L. Network-based in silico drug efficacy screening. *Nat Commun*. 2016;7: 10331. doi:10.1038/ncomms10331
88. Himmelstein DS, Lizee A, Hessler C, Brueggeman L, Chen SL, Hadley D, et al. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *Elife*. 2017;6. doi:10.7554/eLife.26726
89. Greene CS, Krishnan A, Wong AK, Ricciotti E, Zelaya RA, Himmelstein DS, et al. Understanding multicellular function and disease with human tissue-specific networks. *Nat Genet*. 2015;47: 569–576. doi:10.1038/ng.3259
90. Iorio F, Saez-Rodriguez J, di Bernardo D. Network based elucidation of drug response: from modulators to targets. *BMC Syst Biol*. 2013;7: 139. doi:10.1186/1752-0509-7-139
91. Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabási A-L. The human disease network. *Proc Natl Acad Sci U S A*. 2007;104: 8685–8690. doi:10.1073/pnas.0701361104
92. Li Y, Agarwal P. A pathway-based view of human diseases and disease relationships. *PLoS One*. 2009;4: e4346. doi:10.1371/journal.pone.0004346
93. Yang L, Agarwal P. Systematic drug repositioning based on clinical side-effects. *PLoS One*. 2011;6: e28025. doi:10.1371/journal.pone.0028025
94. Ye H, Liu Q, Wei J. Construction of drug network based on side effects and its application for drug repositioning. *PLoS One*. 2014;9: e87864. doi:10.1371/journal.pone.0087864
95. Pan X, Lin X, Cao D, Zeng X, Yu PS, He L, et al. Deep learning for drug repurposing: Methods, databases, and applications. *Wiley Interdiscip Rev*

Comput Mol Sci. 2022;12. doi:10.1002/wcms.1597

96. Aliper A, Plis S, Artemov A, Ulloa A, Mamoshina P, Zhavoronkov A. Deep Learning Applications for Predicting Pharmacological Properties of Drugs and Drug Repurposing Using Transcriptomic Data. *Mol Pharm.* 2016;13: 2524–2530. doi:10.1021/acs.molpharmaceut.6b00248
97. D'Souza S, Prema KV, Balaji S. Machine learning models for drug–target interactions: current knowledge and future directions. *Drug Discov Today.* 2020;25: 748–756. doi:10.1016/j.drudis.2020.03.003
98. Schmitz U, Wolkenhauer O, editors. *Systems Medicine*. 1st ed. New York, NY: Humana Press; 2015. doi:10.1007/978-1-4939-3283-2
99. Bruggeman FJ, Westerhoff HV. The nature of systems biology. *Trends Microbiol.* 2007;15: 45–50. doi:10.1016/j.tim.2006.11.003
100. Auffray C, Chen Z, Hood L. Systems medicine: the future of medical genomics and healthcare. *Genome Med.* 2009;1: 2. doi:10.1186/gm2
101. Barabási A-L. Network medicine--from obesity to the “diseasome.” *The New England journal of medicine.* 2007. pp. 404–407. doi:10.1056/NEJMe078114
102. Barabási A-L, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. *Nat Rev Genet.* 2011;12: 56–68. doi:10.1038/nrg2918
103. Chao S-Y. Graph theory and analysis of biological data in computational biology. *Advanced Technologies.* InTech; 2009. doi:10.5772/8205
104. Riaz F, Ali KM. Applications of graph theory in computer science. 2011 Third International Conference on Computational Intelligence, Communication Systems and Networks. *IEEE*; 2011. doi:10.1109/cicsyn.2011.40
105. Li MM, Huang K, Zitnik M. Graph representation learning in biomedicine and healthcare. *Nat Biomed Eng.* 2022;6: 1353–1369. doi:10.1038/s41551-022-00942-x
106. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 2003;13: 2498–2504. doi:10.1101/gr.1239303
107. Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, Ronzano F, Centeno E, Sanz F, et al. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.* 2020;48: D845–D855. doi:10.1093/nar/gkz1021
108. Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, et al. The STRING database in 2021: customizable protein–protein networks, and functional characterization of user–uploaded gene/measurement sets. *Nucleic Acids Res.* 2021;49: D605–D612. doi:10.1093/nar/gkaa1074

109. Gillespie M, Jassal B, Stephan R, Milacic M, Rothfels K, Senff-Ribeiro A, et al. The reactome pathway knowledgebase 2022. *Nucleic Acids Res.* 2022;50: D687–D692. doi:10.1093/nar/gkab1028
110. Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res.* 2023;51: D587–D592. doi:10.1093/nar/gkac963
111. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* 2019;47: D330–D338. doi:10.1093/nar/gky1055
112. Vasilevsky NA, Matentzoglou NA, Toro S, Flack JE IV, Hegde H, Unni DR, et al. Mondo: Unifying diseases for the world, by the world. *bioRxiv.* 2022. doi:10.1101/2022.04.13.22273750
113. Hogan A, Blomqvist E, Cochez M, D'amato C, Melo GD, Gutierrez C, et al. Knowledge graphs. *ACM Comput Surv.* 2022;54: 1–37. doi:10.1145/3447772
114. Sun Y, Han J. Mining heterogeneous information networks: Principles and methodologies. *Synth Lect Data Min Knowl Discov.* 2012;3: 1–159. doi:10.2200/s00433ed1v01y201207dmk005
115. Stark C, Breitkreutz B-J, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* 2006;34: D535–9. doi:10.1093/nar/gkj109
116. Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, Kerrien S, Orchard S, et al. IntAct: an open source molecular interaction database. *Nucleic Acids Res.* 2004;32: D452–5. doi:10.1093/nar/gkh052
117. Goel R, Harsha HC, Pandey A, Prasad TSK. Human Protein Reference Database and Human Proteinpedia as resources for phosphoproteome analysis. *Mol Biosyst.* 2012;8: 453–463. doi:10.1039/c1mb05340j
118. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 2015;43: D447–52. doi:10.1093/nar/gku1003
119. Alanis-Lobato G, Andrade-Navarro MA, Schaefer MH. HIPPIE v2.0: enhancing meaningfulness and reliability of protein-protein interaction networks. *Nucleic Acids Res.* 2017;45: D408–D414. doi:10.1093/nar/gkw985
120. Barabási A-L, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat Rev Genet.* 2004;5: 101–113. doi:10.1038/nrg1272
121. Goh K-I, Oh E, Jeong H, Kahng B, Kim D. Classification of scale-free networks. *Proc Natl Acad Sci U S A.* 2002;99: 12583–12588. doi:10.1073/pnas.202301299
122. Jeong H, Mason SP, Barabási AL, Oltvai ZN. Lethality and centrality in protein networks. *Nature.* 2001;411: 41–42. doi:10.1038/35075138

123. Broido AD, Clauset A. Scale-free networks are rare. *Nat Commun.* 2019;10: 1017. doi:10.1038/s41467-019-08746-5
124. Lima-Mendez G, van Helden J. The powerful law of the power law and other myths in network biology. *Mol Biosyst.* 2009;5: 1482–1493. doi:10.1039/b908681a
125. Przulj N. Biological network comparison using graphlet degree distribution. *Bioinformatics.* 2007;23: e177–83. doi:10.1093/bioinformatics/btl301
126. Khanin R, Wit E. How scale-free are biological networks. *J Comput Biol.* 2006;13: 810–818. doi:10.1089/cmb.2006.13.810
127. Hahn MW, Kern AD. Comparative genomics of centrality and essentiality in three eukaryotic protein–interaction networks. *Mol Biol Evol.* 2005;22: 803–806. doi:10.1093/molbev/msi072
128. Perfetto L, Pastrello C, Del-Toro N, Duesbury M, Iannuccelli M, Kotlyar M, et al. The IMEx coronavirus interactome: an evolving map of Coronaviridae–host molecular interactions. *Database .* 2020;2020. doi:10.1093/database/baaa096
129. Del Toro N, Shrivastava A, Ragueneau E, Meldal B, Combe C, Barrera E, et al. The IntAct database: efficient access to fine-grained molecular interaction data. *Nucleic Acids Res.* 2022;50: D648–D653. doi:10.1093/nar/gkab1006
130. Lim J, Hao T, Shaw C, Patel AJ, Szabó G, Rual J-F, et al. A protein–protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. *Cell.* 2006;125: 801–814. doi:10.1016/j.cell.2006.03.032
131. Sadegh S, Skelton J, Anastasi E, Maier A, Adamowicz K, Möller A, et al. Lacking mechanistic disease definitions and corresponding association data hamper progress in network medicine and beyond. *Nat Commun.* 2023;14: 1662. doi:10.1038/s41467-023-37349-4
132. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 2019;47: D506–D515. doi:10.1093/nar/gky1049
133. Davis AP, Grondin CJ, Johnson RJ, Sciaky D, Wieggers J, Wieggers TC, et al. Comparative Toxicogenomics Database (CTD): update 2021. *Nucleic Acids Res.* 2021;49: D1138–D1143. doi:10.1093/nar/gkaa891
134. Rehm HL, Berg JS, Brooks LD, Bustamante CD, Evans JP, Landrum MJ, et al. ClinGen – the Clinical Genome Resource. *N Engl J Med.* 2015;372: 2235–2242. doi:10.1056/NEJMs1406261
135. Martin AR, Williams E, Foulger RE, Leigh S, Daugherty LC, Niblock O, et al. PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nat Genet.* 2019;51: 1560–1565. doi:10.1038/s41588-019-0528-2
136. Tamborero D, Rubio-Perez C, Deu-Pons J, Schroeder MP, Vivancos A, Rovira A, et al. Cancer Genome Interpreter annotates the biological and

- clinical relevance of tumor alterations. *Genome Med.* 2018;10: 25.
doi:10.1186/s13073-018-0531-8
137. Gutiérrez-Sacristán A, Grosdidier S, Valverde O, Torrens M, Bravo À, Piñero J, et al. PsyGeNET: a knowledge platform on psychiatric disorders and their genes. *Bioinformatics.* 2015;31: 3075–3077.
doi:10.1093/bioinformatics/btv301
138. Piñero J, Queralt-Rosinach N, Bravo À, Deu-Pons J, Bauer-Mehren A, Baron M, et al. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database .* 2015;2015: bav028.
doi:10.1093/database/bav028
139. Hidalgo CA, Blumm N, Barabási A-L, Christakis NA. A dynamic network approach for the study of human phenotypes. *PLoS Comput Biol.* 2009;5: e1000353. doi:10.1371/journal.pcbi.1000353
140. Folino F, Pizzuti C. Link prediction approaches for disease networks. *Information Technology in Bio- and Medical Informatics.* Berlin, Heidelberg: Springer Berlin Heidelberg; 2012. pp. 99–108. doi:10.1007/978-3-642-32395-9_8
141. Nøhr C, Parv L, Kink P, Cummings E, Almond H, Nørgaard JR, et al. Nationwide citizen access to their health data: analysing and comparing experiences in Denmark, Estonia and Australia. *BMC Health Serv Res.* 2017;17: 534. doi:10.1186/s12913-017-2482-y
142. Sheikh A, Cornford T, Barber N, Avery A, Takian A, Lichtner V, et al. Implementation and adoption of nationwide electronic health records in secondary care in England: final qualitative results from prospective national evaluation in “early adopter” hospitals. *BMJ.* 2011;343: d6054.
doi:10.1136/bmj.d6054
143. Schmitt T. Implementing electronic Health records in Germany: Lessons (yet to be) learned. *Int J Integr Care.* 2023;23: 13. doi:10.5334/ijic.6578
144. Fotouhi B, Momeni N, Riolo MA, Buckeridge DL. Statistical methods for constructing disease comorbidity networks from longitudinal inpatient data. *Appl Netw Sci.* 2018;3: 46. doi:10.1007/s41109-018-0101-4
145. Siggaard T, Reguant R, Jørgensen IF, Haue AD, Lademann M, Aguayo-Orozco A, et al. Disease trajectory browser for exploring temporal, population-wide disease progression patterns in 7.2 million Danish patients. *Nat Commun.* 2020;11: 4952. doi:10.1038/s41467-020-18682-4
146. Westergaard D, Moseley P, Sørup FKH, Baldi P, Brunak S. Population-wide analysis of differences in disease progression patterns in men and women. *Nat Commun.* 2019;10: 666. doi:10.1038/s41467-019-08475-9
147. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.* 2000;25: 25–29. doi:10.1038/75556

148. Kotlyar M, Pastrello C, Malik Z, Jurisica I. IID 2018 update: context-specific physical protein-protein interactions in human, model organisms and domesticated species. *Nucleic Acids Res.* 2019;47: D581–D589. doi:10.1093/nar/gky1037
149. Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput Biol.* 2012;8: e1002375. doi:10.1371/journal.pcbi.1002375
150. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A.* 2005;102: 15545–15550. doi:10.1073/pnas.0506580102
151. Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, Loscalzo J, et al. Disease networks. Uncovering disease-disease relationships through the incomplete interactome. *Science.* 2015;347: 1257601. doi:10.1126/science.1257601
152. Batra R, Alcaraz N, Gitzhofer K, Pauling J, Ditzel HJ, Hellmuth M, et al. On the performance of de novo pathway enrichment. *NPJ Syst Biol Appl.* 2017;3: 6. doi:10.1038/s41540-017-0007-2
153. Ideker T, Ozier O, Schwikowski B, Siegel AF. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics.* 2002;18 Suppl 1: S233–40. doi:10.1093/bioinformatics/18.suppl_1.s233
154. Guo Z, Wang L, Li Y, Gong X, Yao C, Ma W, et al. Edge-based scoring and searching method for identifying condition-responsive protein-protein interaction sub-network. *Bioinformatics.* 2007;23: 2121–2128. doi:10.1093/bioinformatics/btm294
155. Dreyfus SE, Wagner RA. The steiner problem in graphs. *Networks.* 1971;1: 195–207. doi:10.1002/net.3230010302
156. Leiserson MDM, Vandin F, Wu H-T, Dobson JR, Eldridge JV, Thomas JL, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat Genet.* 2015;47: 106–114. doi:10.1038/ng.3168
157. Paull EO, Carlin DE, Niepel M, Sorger PK, Haussler D, Stuart JM. Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE). *Bioinformatics.* 2013;29: 2757–2764. doi:10.1093/bioinformatics/btt471
158. Komurov K, Dursun S, Erdin S, Ram PT. NetWalker: a contextual network analysis tool for functional genomics. *BMC Genomics.* 2012;13: 282. doi:10.1186/1471-2164-13-282
159. Reyna MA, Leiserson MDM, Raphael BJ. Hierarchical HotNet: identifying hierarchies of altered subnetworks. *Bioinformatics.* 2018;34: i972–i980. doi:10.1093/bioinformatics/bty613

160. Nacu S, Critchley-Thorne R, Lee P, Holmes S. Gene expression network analysis and applications to immunology. *Bioinformatics*. 2007;23: 850–858. doi:10.1093/bioinformatics/btm019
161. Breitling R, Amtmann A, Herzyk P. Graph-based iterative Group Analysis enhances microarray interpretation. *BMC Bioinformatics*. 2004;5: 100. doi:10.1186/1471-2105-5-100
162. Ulitsky I, Karp RM, Shamir R. Detecting disease-specific dysregulated pathways via analysis of clinical expression profiles. *Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2008. pp. 347–359. doi:10.1007/978-3-540-78839-3_30
163. Alcaraz N, Küçük H, Weile J, Wipat A, Baumbach J. KeyPathwayMiner: Detecting case-specific biological pathways using expression data. *Internet Math*. 2011;7: 299–313. doi:10.1080/15427951.2011.604548
164. Gu J, Chen Y, Li S, Li Y. Identification of responsive gene modules by network-based gene clustering and extending: application to inflammation and angiogenesis. *BMC Syst Biol*. 2010;4: 47. doi:10.1186/1752-0509-4-47
165. Wu G, Stein L. A network module-based method for identifying cancer prognostic signatures. *Genome Biol*. 2012;13: R112. doi:10.1186/gb-2012-13-12-r112
166. Casas AI, Hassan AA, Larsen SJ, Gomez-Rangel V, Elbatreek M, Kleikers PWM, et al. From single drug targets to synergistic network pharmacology in ischemic stroke. *Proc Natl Acad Sci U S A*. 2019;116: 7129–7136. doi:10.1073/pnas.1820799116
167. Li J, Zheng S, Chen B, Butte AJ, Swamidass SJ, Lu Z. A survey of current trends in computational drug repositioning. *Brief Bioinform*. 2016;17: 2–12. doi:10.1093/bib/bbv020
168. Nogales C, Mamdouh ZM, List M, Kiel C, Casas AI, Schmidt HHHW. Network pharmacology: curing causal mechanisms instead of treating symptoms. *Trends Pharmacol Sci*. 2022;43: 136–150. doi:10.1016/j.tips.2021.11.004
169. Hopkins AL. Network pharmacology: the next paradigm in drug discovery. *Nat Chem Biol*. 2008;4: 682–690. doi:10.1038/nchembio.118
170. Schriml LM, Arze C, Nadendla S, Chang Y-WW, Mazaitis M, Felix V, et al. Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Res*. 2012;40: D940–6. doi:10.1093/nar/gkr972
171. Schriml LM, Munro JB, Schor M, Olley D, McCracken C, Felix V, et al. The Human Disease Ontology 2022 update. *Nucleic Acids Res*. 2022;50: D1255–D1261. doi:10.1093/nar/gkab1063
172. World Health Organization. The International Statistical Classification of Diseases and Health Related Problems ICD-10: Tenth Revision. Volume 2:

Instruction Manual. World Health Organization; 2004. Available: https://books.google.com/books/about/The_International_Statistical_Classifica.html?hl=&id=u3KSRLclIOoC

173. National Library of Medicine (U.S.). Medical Subject Headings: Main Headings, Subheadings and Cross References Used in the Index Medicus and the National Library of Medicine Catalog. 1960. Available: <https://play.google.com/store/books/details?id=WfWEqzqGpQEC>
174. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* 2004;32: D267–70. doi:10.1093/nar/gkh061
175. Lee D, de Keizer N, Lau F, Cornet R. Literature review of SNOMED CT use. *J Am Med Inform Assoc.* 2014;21: e11–9. doi:10.1136/amiajnl-2013-001636
176. Amberger JS, Bocchini CA, Scott AF, Hamosh A. OMIM.org: leveraging knowledge across phenotype–gene relationships. *Nucleic Acids Res.* 2019;47: D1038–D1043. doi:10.1093/nar/gky1151
177. Schmidt HHH. The end of medicine as we know it – and why your health has a future. Springer Nature; 2022. Available: <https://play.google.com/store/books/details?id=khZuEAAAQBAJ>
178. Agache I, Akdis CA. Precision medicine and phenotypes, endotypes, genotypes, regiotypes, and theratypes of allergic diseases. *J Clin Invest.* 2019;129: 1493–1503. doi:10.1172/JCI124611
179. Loscalzo J, Kohane I, Barabasi A-L. Human disease classification in the postgenomic era: a complex systems approach to human pathobiology. *Mol Syst Biol.* 2007;3: 124. doi:10.1038/msb4100163
180. Anderson GP. Endotyping asthma: new insights into key pathogenic mechanisms in a complex, heterogeneous disease. *Lancet.* 2008;372: 1107–1119. doi:10.1016/S0140-6736(08)61452-X
181. Lötvall J, Akdis CA, Bacharier LB, Bjermer L, Casale TB, Custovic A, et al. Asthma endotypes: a new approach to classification of disease entities within the asthma syndrome. *J Allergy Clin Immunol.* 2011;127: 355–360. doi:10.1016/j.jaci.2010.11.037
182. Searls DB. Data integration: challenges for drug discovery. *Nat Rev Drug Discov.* 2005;4: 45–58. doi:10.1038/nrd1608
183. Timón-Reina S, Rincón M, Martínez-Tomás R. An overview of graph databases and their applications in the biomedical domain. *Database.* 2021;2021. doi:10.1093/database/baab026
184. Köhler J, Baumbach J, Taubert J, Specht M, Skusa A, Rüegg A, et al. Graph-based analysis and visualization of experimental results with ONDEX. *Bioinformatics.* 2006;22: 1383–1390. doi:10.1093/bioinformatics/btl081
185. Bastian M, Heymann S, Jacomy M. Gephi: An open source software for

- exploring and manipulating networks. Proceedings of the International AAAI Conference on Web and Social Media. 2009;3: 361–362.
doi:10.1609/icwsm.v3i1.13937
186. Have CT, Jensen LJ. Are graph databases ready for bioinformatics? *Bioinformatics*. 2013;29: 3107–3108. doi:10.1093/bioinformatics/btt549
 187. Rodriguez MA, Neubauer P. The graph traversal pattern. 2010.
doi:10.48550/ARXIV.1004.1001
 188. Francis N, Green A, Guagliardo P, Libkin L, Lindaaker T, Marsault V, et al. Cypher. Proceedings of the 2018 International Conference on Management of Data. New York, NY, USA: ACM; 2018. doi:10.1145/3183713.3190657
 189. Fensel D, Simsek U, Angele K, Huaman E, Karle E, Panasiuk O, et al. Knowledge graphs: Methodology, tools and selected use cases. 1st ed. Cham, Switzerland: Springer Nature; 2020. doi:10.1007/978-3-030-37439-6
 190. Jose B, Abraham S. Exploring the merits of nosql: A study based on mongodb. 2017 International Conference on Networks & Advances in Computational Technologies (NetACT). IEEE; 2017.
doi:10.1109/netact.2017.8076778
 191. Jose B, Abraham S. Performance analysis of NoSQL and relational databases with MongoDB and MySQL. *Mater Today*. 2020;24: 2036–2043.
doi:10.1016/j.matpr.2020.03.634
 192. Avram S, Bologna CG, Holmes J, Bocci G, Wilson TB, Nguyen D-T, et al. DrugCentral 2021 supports drug discovery and repositioning. *Nucleic Acids Res*. 2021;49: D1160–D1169. doi:10.1093/nar/gkaa997
 193. Kacprowski T, Doncheva NT, Albrecht M. NetworkPrioritizer: a versatile tool for network-based prioritization of candidate disease genes or other molecules. *Bioinformatics*. 2013;29: 1471–1473.
doi:10.1093/bioinformatics/btt164
 194. Alcaraz N, List M, Dissing-Hansen M, Rehmsmeier M, Tan Q, Mollenhauer J, et al. Robust de novo pathway enrichment with KeyPathwayMiner 5. *F1000Res*. 2016;5: 1531. doi:10.12688/f1000research.9054.1
 195. Gyöngyi Z, Garcia-Molina H, Pedersen J. Combating Web Spam with TrustRank. Proceedings 2004 VLDB Conference. 2004. pp. 576–587.
doi:10.1016/b978-012088469-8.50052-8
 196. Ghiassian SD, Menche J, Barabási A-L. A DIseAse MOdule Detection (DIAMOND) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. *PLoS Comput Biol*. 2015;11: e1004120. doi:10.1371/journal.pcbi.1004120
 197. Kou L, Markowsky G, Berman L. A fast algorithm for Steiner trees. *Acta Inform*. 1981;15: 141–145. doi:10.1007/BF00288961
 198. Järvelin K, Kekäläinen J. Cumulated gain-based evaluation of IR

- techniques. *ACM Trans Inf Syst Secur.* 2002;20: 422–446.
doi:10.1145/582415.582418
199. Sanfeliu A, Fu K-S. A distance measure between attributed relational graphs for pattern recognition. *IEEE Trans Syst Man Cybern.* 1983;SMC-13: 353–362. doi:10.1109/tsmc.1983.6313167
200. *Systems Medicine: Integrative, Qualitative and Computational Approaches.* Academic Press; 2020. Available: <https://play.google.com/store/books/details?id=suneDwAAQBAJ>
201. Sun W, Sanderson PE, Zheng W. Drug combination therapy increases successful drug repositioning. *Drug Discov Today.* 2016;21: 1189–1195. doi:10.1016/j.drudis.2016.05.015
202. Liu H, Zhang W, Nie L, Ding X, Luo J, Zou L. Predicting effective drug combinations using gradient tree boosting based on features extracted from drug-protein heterogeneous network. *BMC Bioinformatics.* 2019;20: 645. doi:10.1186/s12859-019-3288-1
203. Brimacombe KR, Zhao T, Eastman RT, Hu X, Wang K, Backus M, et al. An OpenData portal to share COVID-19 drug repurposing data in real time. *bioRxiv.* 2020. doi:10.1101/2020.06.04.135046
204. Janes J, Young ME, Chen E, Rogers NH, Burgstaller-Muehlbacher S, Hughes LD, et al. The ReFRAME library as a comprehensive drug repurposing library and its application to the treatment of cryptosporidiosis. *Proc Natl Acad Sci U S A.* 2018;115: 10750–10755. doi:10.1073/pnas.1810137115
205. Jarada TN, Rokne JG, Alhadj R. A review of computational drug repositioning: strategies, approaches, opportunities, challenges, and directions. *J Cheminform.* 2020;12: 46. doi:10.1186/s13321-020-00450-7
206. Morselli Gysi D, do Valle Í, Zitnik M, Ameli A, Gan X, Varol O, et al. Network medicine framework for identifying drug-repurposing opportunities for COVID-19. *Proc Natl Acad Sci U S A.* 2021;118. doi:10.1073/pnas.2025581118
207. Seyhan AA. Lost in translation: the valley of death across preclinical and clinical divide – identification of problems and overcoming obstacles. *Transl Med Commun.* 2019;4. doi:10.1186/s41231-019-0050-7
208. Jinawath N, Bunbanjerdsuk S, Chayanupatkul M, Ngamphaiboon N, Asavapanumas N, Svasti J, et al. Bridging the gap between clinicians and systems biologists: from network biology to translational biomedical research. *J Transl Med.* 2016;14: 324. doi:10.1186/s12967-016-1078-3
209. AbdulHameed MDM, Tawa GJ, Kumar K, Ippolito DL, Lewis JA, Stallings JD, et al. Systems level analysis and identification of pathways and networks associated with liver fibrosis. *PLoS One.* 2014;9: e112193. doi:10.1371/journal.pone.0112193
210. Sharma A, Halu A, Decano JL, Padi M, Liu Y-Y, Prasad RB, et al.

Controllability in an islet specific regulatory network identifies the transcriptional factor NFATC4, which regulates Type 2 Diabetes associated genes. *NPJ Syst Biol Appl.* 2018;4: 25. doi:10.1038/s41540-018-0057-0

211. Maron BA, Wang R-S, Shevtsov S, Drakos SG, Arons E, Wever-Pinzon O, et al. Individualized interactomes for network-based precision medicine in hypertrophic cardiomyopathy with implications for other clinical pathophenotypes. *Nat Commun.* 2021;12: 873. doi:10.1038/s41467-021-21146-y
212. Halu A, Liu S, Baek SH, Hobbs BD, Hunninghake GM, Cho MH, et al. Exploring the cross-phenotype network region of disease modules reveals concordant and discordant pathways between chronic obstructive pulmonary disease and idiopathic pulmonary fibrosis. *Hum Mol Genet.* 2019;28: 2352–2364. doi:10.1093/hmg/ddz069
213. Larsen SJ, Schmidt HHHW, Baumbach J. De Novo and supervised endophenotyping using network-guided ensemble learning. *Syst Med.* 2020;3: 8–21. doi:10.1089/sysm.2019.0008
214. Zolotareva O, Khakabimamaghani S, Isaeva OI, Chervontseva Z, Savchik A, Ester M. Identification of differentially expressed gene modules in heterogeneous diseases. *Bioinformatics.* 2021;37: 1691–1698. doi:10.1093/bioinformatics/btaa1038
215. Levi H, Elkon R, Shamir R. DOMINO: a network-based active module identification algorithm with reduced rate of false calls. *Mol Syst Biol.* 2021;17: e9593. doi:10.15252/msb.20209593
216. Schaefer MH, Serrano L, Andrade-Navarro MA. Correcting for the study bias associated with protein-protein interaction measurements reveals differences between protein degree distributions from different cancer types. *Front Genet.* 2015;6: 260. doi:10.3389/fgene.2015.00260
217. Wachi S, Yoneda K, Wu R. Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. *Bioinformatics.* 2005;21: 4205–4208. doi:10.1093/bioinformatics/bti688
218. Jonsson PF, Bates PA. Global topological features of cancer proteins in the human interactome. *Bioinformatics.* 2006;22: 2291–2297. doi:10.1093/bioinformatics/btl390
219. Rambaldi D, Giorgi FM, Capuani F, Ciliberto A, Ciccarelli FD. Low duplicability and network fragility of cancer genes. *Trends Genet.* 2008;24: 427–430. doi:10.1016/j.tig.2008.06.003
220. Dunham I. Human genes: Time to follow the roads less traveled? *PLoS Biol.* 2018;16: e3000034. doi:10.1371/journal.pbio.3000034
221. Stoeger T, Gerlach M, Morimoto RI, Nunes Amaral LA. Large-scale investigation of the reasons why potentially important genes are ignored. *PLoS Biol.* 2018;16: e2006643. doi:10.1371/journal.pbio.2006643

222. Haynes WA, Tomczak A, Khatri P. Gene annotation bias impedes biomedical research. *Sci Rep.* 2018;8: 1362. doi:10.1038/s41598-018-19333-x
223. Lazareva O, Baumbach J, List M, Blumenthal DB. On the limits of active module identification. *Brief Bioinform.* 2021;22. doi:10.1093/bib/bbab066
224. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. *Nature.* 2021;596: 583–589. doi:10.1038/s41586-021-03819-2
225. Brown AS, Patel CJ. A review of validation strategies for computational drug repositioning. *Brief Bioinform.* 2018;19: 174–177. doi:10.1093/bib/bbw110
226. Brown AS, Patel CJ. A standard database for drug repositioning. *Sci Data.* 2017;4: 170029. doi:10.1038/sdata.2017.29
227. Maier A, Hartung M, Abovsky M, Adamowicz K, Bader GD, Baier S, et al. Drugst.One -- A plug-and-play solution for online systems medicine and network-based drug repurposing. 2023. doi:10.48550/ARXIV.2305.15453