

ALISA PAVEL

**Knowledge Graphs,  
Network Models and  
Health Data Science  
Approaches for  
Toxicology and  
Pharmacology**



ALISA PAVEL

Knowledge Graphs, Network Models and  
Health Data Science Approaches  
for Toxicology and Pharmacology

ACADEMIC DISSERTATION

To be presented, with the permission of  
the Faculty of Medicine and Health Technology  
of Tampere University,  
for public discussion in the Jarmo Visakorpi Sali  
of the Kauppi Campus, Arvo Ylpön katu 34, 33520 Tampere,  
on 12 April 2024, at 11 o'clock.

## ACADEMIC DISSERTATION

Tampere University, Faculty of Medicine and Health Technology  
Finland

*Responsible  
supervisor  
and Custos*

Professor Dario Greco  
Tampere University  
Finland

*Pre-examiners*

Professor Andrea Ganna  
University of Helsinki  
Finland

PhD Thomas Exner  
The Technical University of Darmstadt  
Germany

*Opponent*

Professor Jacques Fleuriot  
The University of Edinburgh  
UK

The originality of this thesis has been checked using the Turnitin OriginalityCheck service.

Copyright ©2024 author

Cover design: Roihu Inc.

ISBN 978-952-03-3342-3 (print)

ISBN 978-952-03-3343-0 (pdf)

ISSN 2489-9860 (print)

ISSN 2490-0028 (pdf)

<http://urn.fi/URN:ISBN:978-952-03-3343-0>



Carbon dioxide emissions from printing Tampere University dissertations have been compensated.

PunaMusta Oy – Yliopistopaino  
Joensuu 2024

# ACKNOWLEDGEMENT

The thesis work was conducted at the Finnish Hub for Development and Validation of Integrated Approaches (FHAIIVE) at Tampere University, Faculty of Medicine and Health Technology.

I want to thank everyone who supported me during this time. Special thanks to my supervisor Prof. Dario Greco for giving me the opportunity to implement my vision of a large-scale Knowledge Graph. I want to thank PhD Angela Serra for her continuous support and discussions, and every other current or past member of FHAIIVE who made this possible, especially the PhD students that started their journey with me.

Special thanks to Professor Ian Simpson (The University of Edinburgh), not only for being a member of my thesis committee but also your support during my masters and encouragement to pursue a PhD in Bioinformatics. I am grateful to Assistant Professor Vittorio Fortino (University of Eastern Finland) for acting as a member of my thesis committee.

Big thanks to my family for their support. I am especially grateful to J. for supporting me all this time and happily moving with me to wherever life takes me. And finally, I want to thank all the homeless cats of HML, with honorary mentions of P. & N., then who has time to worry about a thesis if they have you to worry about.



# ABSTRACT

Big Data analytics focus on the collection, modelling, and analysis of large-scale data to identify correlations and relationships, gain new insights as well as to make predictions about possible future outcomes or new facts about the world under investigation. In the health sciences, Big Data can have many different facets and applications, ranging from hospital process optimization, over image classification and personalized medicine to drug development and chemicals safety assessment. However, current Big Data studies in the life sciences are often limited to a small range of data sources and data types due to the diversity and complexity of the available data, standards, and interpretations.

Knowledge Graphs are a highly flexible, link-oriented data structure, which, based on the application of a reasoning engine, allow the inference of new facts about the world under investigation. Knowledge Graphs are built upon a graph data model, which is a schema-free, highly flexible, and modifiable data management model. In addition to classical data retrieval and analytical methodologies, graph-based data models are link and path focused as well as allow the application of network metrics to analyse not only individual data points, but with respect to the whole system.

In this thesis I have investigated the use of graph data models and Knowledge Graphs as data management, data integration and knowledge inference engines for the highly diverse data across the life sciences with a focus on their application to the compound safety and development process. In addition, I developed and collected different network analysis methodologies for the analysis of networks created from molecular data as well as networks contained directly or indirectly in a Knowledge Graph data model or in combination with molecular data.





# CONTENTS

Acknowledgement.....	iii
Abstract .....	v
Original Publications .....	xix
Author Contribution.....	xx
1 Introduction.....	21
2 Health Data Science .....	24
2.1 Data Science and Big Data Analytics .....	24
2.2 Large Scale Data in the Life Sciences.....	25
2.2.1 Healthcare .....	26
2.2.2 Personalized Medicine .....	26
2.2.3 Hospital/ Healthcare Optimization .....	28
2.2.4 Image/ Disease Recognition .....	28
2.2.5 Clinical Trials.....	28
2.2.6 In Toxicology, Drug Design & Chemical Safety.....	29
3 Data Integration.....	35
3.1 Horizontal Data Integration.....	36
3.2 Vertical Data Integration.....	37
3.3 Diagonal Data Integration .....	37
3.4 Challenges of Big Data & Data Integration in the Life Sciences .....	37
3.4.1 Entity Identification and Mapping .....	37
3.4.2 Data Design does not take Big Data Applications into Account .....	39
4 Network Analysis in the Life Sciences .....	41
4.1 Network Inference from Transcriptomics.....	41
4.2 Network Metrics and Their Applications in Toxicology and Pharmacology.....	42
5 Knowledge Graphs.....	48
5.1 Advantages of a Graph Data Model .....	48
5.1.1 Leveraging the Knowledge Graph Topology .....	49

5.2	Knowledge Graphs as Predictive Engines.....	50
5.3	Examples of Knowledge Graphs in Toxicology, Chemical Safety and Drug Development.....	51
6	Aims of the Study.....	54
7	Materials and Methods.....	55
7.1	The Unified Knowledge Space .....	55
7.1.1	Technology .....	56
7.1.2	Data Model.....	58
7.1.3	Data Integration.....	62
7.1.4	Data Retrieval.....	64
7.2	Intermediate Genes – Identifying Relevant Non-Measured Genes .....	70
7.3	VOLTA – Network Analytics & Multi Network Clustering.....	71
7.3.1	Comparison of Drug MOA via Network Analysis.....	72
7.3.2	Grouping Networks Based on Network Metrics.....	73
7.4	Grouping of MOA Across Systems and Data Sets .....	76
7.4.1	Extraction of a Homogenous Multilayer Network.....	76
7.4.2	Detection of Similar Genes.....	77
7.4.3	KNeMAP Vector.....	77
8	Results.....	78
8.1	The UKS as a Robust Multidimensional Data Source for Data Retrieval, Knowledge Linkage and to Support Transcriptomic Analysis .....	78
8.1.1	Robust, Multi-Source Support Data.....	83
8.1.2	Multidimensional Data.....	83
8.1.3	Linking Between Independent Data Sets and Data Points .....	84
8.2	Robust Prior Information Can Reduce Noise in Gene Expression Data .....	85
8.3	Identifying Non-Measured Relevant Genes .....	86
8.4	A Comprehensive Network Analysis and Comparison Library.....	87
8.4.1	Modules.....	88
8.4.2	Integration into a Toxicogenomic-Analysis Pipeline.....	90
8.5	Differential Analysis of Co-Expression Networks.....	91
8.6	Characterizing the Impact of Exposure System on a Compound’s MOA.....	92
8.6.1	Identifying Compounds with Similar MOA Across Biological Systems and Across Data Sets .....	93
8.7	Linking Engineered Nanomaterials and Drugs based on their MOA.....	94
9	Discussion.....	96
10	Summary and Conclusion.....	103

References .....	105
Appendix A – The Unified Knowledge Space.....	162
The UKS – Materials and Methods.....	162
The UKS – Results.....	173

## List of Figures

<b>Figure 1</b>	Knowledge Graphs are data models, where entities and relationships are modelled in a network-based format to which semantic meaning is added. A KG can be used as a database for knowledge retrieval, analysed as a graph and allows the application of a reasoning engine to infer new facts about the world under investigation.....	22
<b>Figure 2</b>	Complexity and diversity of Big Data in healthcare, which can lead to improved patient care and safety through the application of health Data Science.....	25
<b>Figure 3</b>	Vertical data integration, combines data across multiple data types, horizontal data integration combines data of the same type across multiple data sources and diagonal data integration combines data of different types across different data sources.....	36
<b>Figure 4</b>	Node w has the highest degree centrality due to being the node with the most edges, b the highest eigenvector centrality due to being connected to high value nodes, such as w and d, f the highest betweenness centrality due to being the only path between g, i, k and all the other nodes and c the highest closeness centrality due to being closely connected to many of the nodes in the network. Figure adapted from (Pavel, Serra, et al., 2022).....	45
<b>Figure 5</b>	A) graph diameter (red) and B) graph radius (red).....	46
<b>Figure 6</b>	A network with three communities, w, x and z. Figure adapted from (Pavel, Serra, et al., 2022).....	47
<b>Figure 7</b>	Shows the integration of data into a KG and how hidden links can become visible in a network data model. The figure is taken from (Pavel, Saarimäki, et al., 2022).....	50
<b>Figure 8</b>	Describes different types of information and the data layers they can contribute to. Data types only containing entity information (data source X, coloured pink) contain data modelled as nodes and their attributes in the UKS, data sources containing data contributing to the relationship layer (data source Y, coloured green) contain information about the relationships of two entities, which may or may not be of the same type. Lastly data sources can contribute to both layers (data source Z, coloured red and yellow) by containing entity specific information as well as information about the relationships between two entities.....	56

<b>Figure 9</b>	Cypher queries with their graph and textual representation. Figure taken from (Pavel, Saarimäki, et al., 2022) .....	57
<b>Figure 10</b>	Simplified data model of the UKS. In-node labels denote different sub-labels of the main node types, which are displayed in bold outside of the node. ....	60
<b>Figure 11</b>	A) Example of a GENE entity node in the UKS and some of its attributes, which provide further information about the node. B) Example of a P_P_INTERACTION edge in the UKS and some of its attributes. ....	64
<b>Figure 12</b>	An example of a homogenous node and edge network, showcasing the protein protein interaction layer in the UKS. ....	65
<b>Figure 13</b>	Extraction of a robust network from the UKS based on a global threshold. The edge weight indicates its strength of the relationships.....	66
<b>Figure 14</b>	A robust network is extracted based on local (per node) estimated edge weight thresholds. In the example for each node only the edges with at least the maximum edge weight of all the edges connecting to that node are considered. An edge is kept if this is true for at least one of the nodes connected by the edge. The values in the nodes indicate the computed local edge weight threshold for that node. ....	67
<b>Figure 15</b>	Example of a multilayer network in the UKS, combining multiple edge types and node types. Grey) a network of the same node type connected by different edge types. Green) a network connecting different node types by the same edge type. Blue) a network combining different node types connected by different edge types. Entity and relationship types are indicated by their node and edge colour respectively.....	68
<b>Figure 16</b>	Path based data extraction in the UKS. The example showcases a Cypher query stating the node and edge types to be visited to retrieve a possible drug for the treatment of a disease by specifying that the drug target needs to be associated to the disease. ....	69
<b>Figure 17</b>	Shows how intermediate nodes (genes) can be identified on a network via shortest paths. A) Node set 1, can for example be genes known to directly interact with a compound. B) Node set 2, can for example be genes measured as differential expressed in an exposure study. C) Nodes on the shortest paths between node set 1 and node set 2, this can for example be genes propagating the signal from the directly targeted genes to the measured genes. These genes can provide further insight into the molecular processes taking place.....	71

<b>Figure 18</b>	Clustering pipeline available in VOLTA. Image taken from publication II.....	74
<b>Figure 19</b>	Dimensionality of data points stored in the UKS. ....	78
<b>Figure 20</b>	How meta-paths (bold) in the UKS are used to infer genes associated with key events. Figure adapted from (Saarimäki, Morikka, et al., 2023). ....	84
<b>Figure 21</b>	UKS sub-network, containing gene (product) nodes, their interactions as well as drug nodes and drug gene target edges. Paths between genes, associated with different stages of a SARS-CoV-2 infection (red, blue, green), targeted by drugs are indicated by bold edges. ....	85
<b>Figure 22</b>	Showcases the impact artificial added noise to gene expression data has on the stability of MOA vectors, computed with KNeMAP, gene deregulation analysis, fold changes and GSEA. KNeMAPs lowest AUC score indicates the least divergence from the baseline. Figure taken from publication III.....	86
<b>Figure 23</b>	Shows the number of genes known to be associated to COVID-19 (at the time of publication) and how it can be increased with network analytics (set of IN genes). Figure taken form publication I.....	87
<b>Figure 24</b>	Showcases the modules and sub-modules implemented in VOLTA. Figure taken from publication II. ....	88
<b>Figure 25</b>	The NEXCAST software suite, combining multiple software needed for comprehensive toxicogenomic analysis, including VOLTA as a network analysis and comparison module. Figure taken from (Serra, Saarimäki, et al., 2022).....	91
<b>Figure 26</b>	The three MOA clusters detected with VOLTA across 20 different cell lines treated with dasatinib. Cluster B contains mainly cell lines derived from breast tissue, cluster C only cell lines derived from healthy tissue and cluster A contains mostly epithelial (like) carcinoma cells.....	93
<b>Figure A 1</b>	Data and System infrastructure used to manage and deploy the data. External data sources are processed and integrated into a graph database storage, hosted on a NAS (Network Attached Storage). Experimental data, which when processed, is not suitable to be hosted in a graph data model is stored in a file storage and its metadata is stored in the graph database storage. The file storage is	

hosted on a NAS, accessible from both computing servers and personal devices, when connected to the VPN (Virtual Private Network). The graph database storage is read and hosted with Neo4j, which is deployed via Docker on the database server. The deployed database management system can be accessed from both computing servers and personal devices when situated within the VPN. Access security is managed by the VPN access, which is due to the completely internal use and limited number of users enough. Also, while some of the data may be licensed, none of the hosted data are of sensitive nature..... 170

## List of Tables

<b>Table 1</b>	Examples of network metrics that can be applied to biological networks.....	44
<b>Table 2</b>	ENM core materials and known drugs with a similar MOA, computed across datasets with publication III KNeMAP.....	95
<b>Table A 1</b>	Data types and sources integrated into the UKS at this point of time. The data is grouped into 5 different categories: a) interaction: describes direct interactions between two nodes, such as protein protein interactions or drug interactions, b) regulation: describes a regulation relationship between two nodes, such as transcription factor gene regulation, c) functional: describes information about a nodes function or its mechanism of action, such as pathways or a compounds effect on the system, d) associations: describe relationships between nodes that are not of the previous three types, such as ontologies, and e) informative: contains additional knowledge about a node, such as different names, identifiers or sub-structures. To date the UKS contains information collected from more than 80 different databases and data collections. The last column describes if the data in the sources mainly contributes to the relationship layer, entity layer or both, as described in Figure 8. ....	162
<b>Table A 2</b>	Custom created indices in the UKS. Indices are only created for nodes, since the default LOOKUP index covers node label and edge type searches and most of the expected UKS queries are node and relationship type focused, meaning edge attribute-based searches are not expected. Due to the more expensive (with respect of storage) nature of TEXT indices, they are only created for data points where substring searchers are expected, such as PHENOTYPE or CHEMICAL names. For ID based attributes equality searchers are expected, hence RANGE indices have been selected. To save storage, indices have only been created for highly queried or node types for which large amounts of data are available. The number of nodes per label are displayed in Table A 3.....	171
<b>Table A 3</b>	The main UKS node types and their selected identification system to which other identification systems are mapped. ....	172
<b>Table A 4</b>	Node labels and to date data point count in the UKS. To date there are ~68 million data points stored on the nodes in the UKS.....	173



<b>Table A 5</b>	Edge types and to date data points stored on the relationships in the UKS. To date there are ~3.3 billion edge data points stored in the UKS. ....	177
<b>Table A 6</b>	To date data points stored in the file storage managed by the UKS. If a data set varies in their number of genes or samples across the collection an estimate is used. ....	183

# ABBREVIATIONS

AI	Artificial Intelligence
AOP	Adverse Outcome Pathways
API	Application Programming Interface
ARACNE	Algorithm for the Reconstruction of Accurate Cellular Networks
AUC	Area Under the Curve
CLR	Context Likelihood or Relatedness
CMap	Connectivity Map
CPU	Central Processing Unit
CTKG	Clinical Trials Knowledge Graph
DI	Data Integration
DFP	Discriminant Fuzzy Pattern
DNA	Deoxyribonucleic Acid
ENM	Engineered Nanomaterial
FAIR	Findable, Accessible, Interoperable and Reproducible
FDA	Food and Drug Administration
GO	Gene Ontology
GPU	Graphics Processing Unit
GSEA	Gene Set Enrichment Analysis
ICD	International Classification of Diseases
ID	Identifier
INFORM	Inference of Network Response Modules
IT	Information Technology
KE	Key Event
KEGG	Kyoto Encyclopedia of Genes and Genomes
KG	Knowledge Graph
	Network Mapping Approach for Knowledge-Driven Comparison
KNeMAP	Transcriptomic Profiles
LINCS	Library of Integrated Network-based Cellular Signature
miRNA	micro-Ribonucleic Acid
MOA	Mechanism of Action
MRI	Magnetic Resonance Imaging

MRNET	Minimum Redundancy Networks
NAS	Network Attached Storage
NCBI	National Center for Biotechnology Information
NLP	Natural Language Processing
OCT	Optical Coherence Tomography
PCR	Polymerase Chain Reaction
PPI	Protein Protein Interaction
QSAR	Quantitative Structure Activity Relationship
RNA	Ribonucleic Acid
SAR	Structure Active Relationship
SMILES	Simplified Molecular Input Line Entry System
SQL	Structured Query Language
UKS	Unified Knowledge Space
VOLTA	Advanced Molecular Network Analysis
VPN	Virtual Private Network
X-ray	X-radiation



## ORIGINAL PUBLICATIONS

- I. Pavel, Alisa\*, Giusy Del Giudice\*, Antonio Federico, Antonio Di Lieto, Pia AS Kinaret, Angela Serra, and Dario Greco. "Integrated network analysis reveals new genes suggesting COVID-19 chronic effects and treatment." *Briefings in bioinformatics* 22, no. 2 (2021): 1430-1441.  
  
\*Contributed equally
  
- II. Pavel, Alisa, Antonio Federico, Giusy Del Giudice, Angela Serra, and Dario Greco. "Volta: adVanced mOLecular neTwork analysis." *Bioinformatics* 37, no. 23 (2021): 4587-4588.
  
- III. Pavel, Alisa, Giusy Del Giudice, Michele Fratello, Leo Ghemtio, Antonio Di Lieto, Jari Yli-Kauhaluoma, Henri Xhaard, Antonio Federico, Angela Serra, and Dario Greco. "KNeMAP: a network mapping approach for knowledge-driven comparison of transcriptomic profiles." *Bioinformatics* 39, no. 6 (2023): btad341.

# AUTHOR CONTRIBUTION

## **Publication I**

I have performed the collection of prior knowledge, created, and conceptualized the Knowledge Graph framework as well as developed and implemented the network-based analysis methodology. Del Giudice, performed the collection of experimental data and the analysis and interpretation of the computed gene sets. This thesis will only discuss the network analytical and Knowledge Graph aspects of the work, del Giudice's thesis will discuss the biological interpretation of the results. Del Giudice and I wrote the manuscript draft, the final manuscript was written by all the co-authors. Pavel and del Giudice have contributed equal to this work.

## **Publication II**

I have conducted the conceptualization, creation and implementation of the whole package, and performed the case studies presented in the paper. I wrote the first draft of the manuscript; the final manuscript contains contributions by all the co-authors.

## **Publication III**

I have performed the conceptualization and implemented of the network-based analysis methods, the collection of prior knowledge and its integration into the Knowledge Graph framework as well as performed the case studies presented in the paper. I wrote the first draft of the manuscript; the final manuscript contains contributions by all the co-authors.

# 1 INTRODUCTION

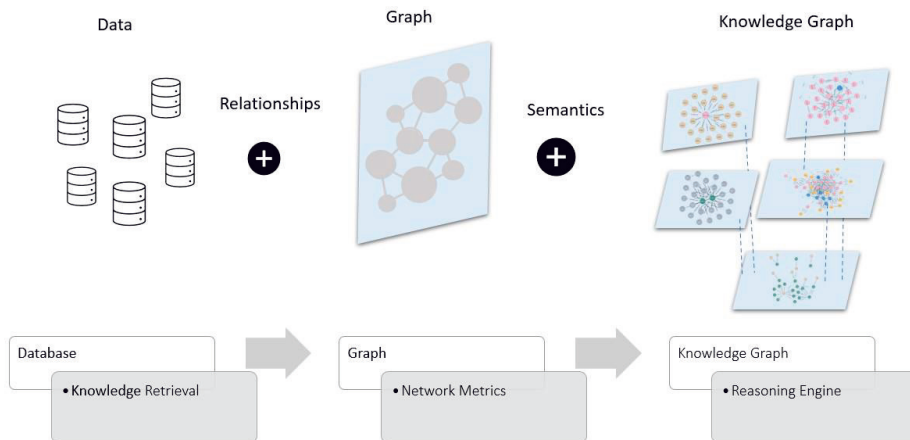
With the constant development of new methodologies and the continuous production of data across different life science sub-domains, the manual processing, analysis, and combination is not feasible. In addition, the chemical industry, develops compounds at a far faster speed than their mechanism of action (MOA) or safety can be evaluated with traditional toxicology methodologies. In the medical field it has been widely acknowledged that the personalised medicine concept, where treatment decisions are based on an individual's behavioural, environmental and genetic factors, is the future (Chang et al., 2018; Cirillo & Valencia, 2019; Fröhlich et al., 2018; Kurnit et al., 2017; Vogenberg et al., 2010). Therefore, there is a need for alternative, computational driven methods, that can handle large scale, highly diverse data sets, produced across the different life science sub-domains (Pavel, Saarimäki, et al., 2022).

Large scale data gathering, and its analysis has become a stable method in many different industries, especially in the information technology (IT) domain, where data is often used to predict user behaviour or needs, to boost sales, advertisement payout or interaction time. Big Data are large collections of high complexity structured and/or unstructured data that due to their size are challenging to analyse and store with traditional data analysis and management methods (Günther et al., 2017; Gupta et al., 2019; Pavel, Saarimäki, et al., 2022; Sagioglu & Sinanc, 2013).

However, large scale Big Data analytics in the life sciences is not as easy as in many other IT-based fields, due to the high variety of data types, data points, reporting quality, standards and even often that the data is not available in computational processable formats (Leonelli, 2019; Pavel, Saarimäki, et al., 2022), which makes automated data integration into a large data model highly challenging. Traditional, relational data models are not suitable due to their non-schema free nature, meaning that a data model needs to be pre-defined, missing data points are not allowed (these will need to be stored as NULL values, increasing the needed storage space unnecessarily) and the data model cannot evolve easily with change in the data. Further joining operations across multiple entity and relationship tables becomes expensive and complex quickly. This limitation suggests that, before the available

life-science data can be used to its full potential, suitable data integration and data modelling solutions, that can handle such complex and diverse data, must be developed or identified. One such schema-free, highly flexible and network driven data model are graph data models, which can be elevated to Knowledge Graphs (KG)s.

Knowledge Graphs are knowledge bases, which model data in a graph-based format, to which a reasoning engine can be applied to infer new facts about the “world under investigation” (Figure 1) (Ehrlinger & Wöß, 2016; Hogan et al., 2021; Pavel, Saarimäki, et al., 2022; Sheth et al., 2019). Entities in the KG are modelled as nodes, while relationships between entities are modelled as edges. By enriching the graph with semantics, additional meaning can be added, which allows reasoning and complex decision making based on the data. Depending on the data model or database management system used, the graph can be directed, undirected, homogeneous, heterogenous, as well as can contain node and edge attributes and/or labels (Hogan et al., 2021; S. Ji et al., 2022). Therefore KGs 1) are knowledge bases, 2) can be analysed in addition to classical data retrieval methods with network-based metrics, and 3) can be used as a reasoning engine (Pavel, Saarimäki, et al., 2022), as displayed in Figure 1.



**Figure 1** Knowledge Graphs are data models, where entities and relationships are modelled in a network-based format to which semantic meaning is added. A KG can be used as a database for knowledge retrieval, analysed as a graph and allows the application of a reasoning engine to infer new facts about the world under investigation.



During my PhD, I have developed the Unified Knowledge Space (UKS), a multi-source, multi-dimension knowledge graph, modelling more than three billion data points across the life science domain, collected from over 80 different independent data sources. This dissertation demonstrates the potential of KGs as a highly flexible and suitable data model for Big Data analytics across the life sciences, as well as showcases that with a combination of computational and manual data integration procedures, large scale multi-dimensional data collection is achievable in the life-sciences. The UKS, to my knowledge, is one of the largest multi-dimensional data sources created in the chemical and drug domain (Abdelaziz et al., 2017; Al-Saleem et al., 2021; Mohamed et al., 2019; Pavel, Saarimäki, et al., 2022; R. Zhang et al., 2021) and can be seen as the data modelling base for further Big Data driven (deep) learning models, which can be applied towards a multitude of different hypothesis, questions and research domains. The UKS has potential as a 1) knowledge base, 2) base for network analytics and 3) knowledge inference engine which is showcased across different case studies from the realm of toxicology.

## 2 HEALTH DATA SCIENCE

### 2.1 Data Science and Big Data Analytics

Large collections of structured and/ or unstructured data, that may come from a vast variety of sources, are of high complexity and therefore are challenging to analyse and store with traditional data analysis and management methods are described as Big Data (Günther et al., 2017; Gupta et al., 2019; Pavel, Saarimäki, et al., 2022; Sagirolu & Sinanc, 2013). The possibly high variety of data types and data sources suggests large varieties in the data with respect to their quality, distribution and creation methodologies. In the field of Data Science and Data Analytics the mining and corresponding analysis of such large collections of data are used to identify correlations, relationships and make predictions about future behaviour by modelling, transforming and analysing the data.

Data Science combines statistical methods with computational methods to gain insight into (large scale) data. This allows for example to identify correlations between data points, optimize business procedures and in general gather knowledge about what is and what could be in the future. Big Data and Data Science find a wide range of applications in the industry, ranging from Amazon<sup>1</sup>, Meta<sup>2</sup> and Google<sup>3</sup> to Process Optimizations and machine part failure predictions. Also in the life sciences, data science and large-scale data are of interest and are becoming increasingly more available. Their application range for example from hospital management, patient care, phenotype classification to drug development and toxicology (Cirillo & Valencia, 2019; Cozzoli et al., 2022; Dash et al., 2019; Leonelli, 2019; Mayo et al., 2017; Pavel, Saarimäki, et al., 2022; Qian et al., 2019; H. Zhu et al., 2014) (Figure 2). Combining Big Data with Data Science can improve insights and predictions by enhancing data quality, robustness and increases the information content. In general, it is considered that prediction accuracy, robustness and generalization power correlate with the amount of (diverse) data available (Gupta et al., 2019; Leonelli,

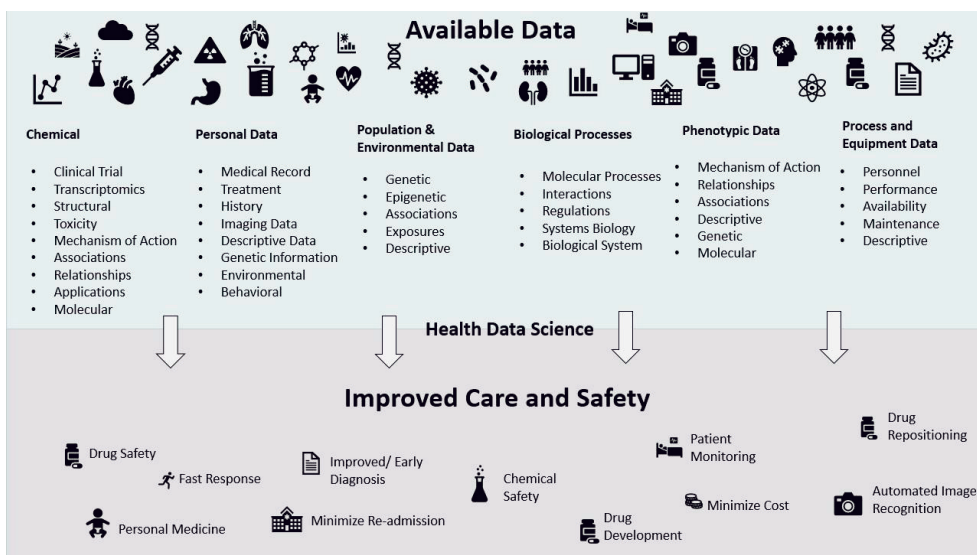
---

<sup>1</sup> amazon.com

<sup>2</sup> meta.com

<sup>3</sup> google.com

2014; Pavel, Saarimäki, et al., 2022) under the constraints of applied methodology, data quality and data availability.



**Figure 2** Complexity and diversity of Big Data in healthcare, which can lead to improved patient care and safety through the application of health Data Science.

## 2.2 Large Scale Data in the Life Sciences

The term Big Data can have different meanings depending on the life science sub-field it is applied to. For example, in chemistry it can refer to large sets of compounds and their structural information, as used in Quantitative Structure Activity Relationship (QSAR) modelling (Lo et al., 2018; Serra et al., 2020), in medicine, Big Data often refers to collections of medical records (Goldstein et al., 2018; Rajkomar et al., 2018), while in molecular biology (such as transcriptomics or proteomics) research it may refer to a large number of samples or exposures (Serra et al., 2020; Tolani et al., 2021; Zielinski et al., 2021) (Figure 2). However, the term “large” is not clearly defined and therefore often refers to “large for that specific field”. This implies that there is a high variation in what is considered high volumes of data, which leads to an expected diversity of Big Data in the life sciences (Pavel, Saarimäki, et al., 2022).

## 2.2.1 Healthcare

Healthcare data can include a wide variety of data, such as clinical information coming from a patient's medical record, vital signs, medical history and other measurable or descriptive data points (Dash et al., 2019; Mehta & Pandit, 2018). In addition, it can contain imaging data, laboratory biomarkers, genetic analysis, -omics data or data coming from medical sensors, such as steps walked in a day or sleeping patterns. This data can further be enriched with clinical trial data, social media data or insurance data (Mehta & Pandit, 2018; Subrahmanya et al., 2022). Integrating and analysing this data allows to generate a complex picture of an individual but also allows them to be put into a broader perspective of other patients and knowledge, potentially allowing to identify currently unknown links or personalized treatment and preventive measures. In addition, healthcare apps and online platforms are becoming more and more available as tools to reduce in person visits as well as to collect patient data, which is made available to a patient's physician or is analysed automatically (Cozzoli et al., 2022). Multiple medical devices, such as glucose measurement devices (Funtanilla et al., 2019; Kamath et al., 2010; Kruger et al., 2019) or heart monitors (Miller et al., 2020) can send data to either a patient's healthcare provider or their own personal mobile devices. This allows the constant monitoring of a patient's condition and possible alert for early changes so that counter measures can be taken as fast as possible, which increases the success rate of performed treatments.

## 2.2.2 Personalized Medicine

Through the increased collection of medical data, such as imaging data, the evolving of technologies, such as whole genome sequencing and Big Data processing platforms, large scale data collection and creation has become of increased interest in personalized healthcare and for the first time makes it a reachable possibility (Cirillo & Valencia, 2019). Personalized medicine aims to shift the focus from a population view to a single patient view, based on the paradigm that a one solution fits all approach neglects the individual needs of patients. It aims at finding the best treatment, or disease management plan for an individual patient instead of the whole population of patients suffering from a given condition. Many diseases (disease processes) display a high heterogeneity, suggesting that homogeneous treatment plans are not suitable but rather individual tuned options should be considered (Goetz & Schork, 2018). It takes factors such as personal environmental conditions, genetics and medical history into account, to identify the most likely successful

treatment with the least adverse outcomes (Vogenberg et al. 2010; Fröhlich et al. 2018; Cirillo and Valencia 2019). Personalized medicine, however, is dependent on the integration and analysis of large numbers of heterogeneous data (sources) (Silva et al., 2022), of which the lack of consistent standards is currently one of the limiting factors. It is widely acknowledged that cancer is a highly heterogeneous disease, even between its known sub-types, and therefore custom treatments are suggested to be more successful (Chang et al., 2018; Dumbrava & Meric-Bernstam, 2018; Fröhlich et al., 2018; Kurnit et al., 2017; Lee et al., 2018).

Electronic health records have a large potential to be mined for their data (Jensen et al., 2012; Rajkomar et al., 2018). On one hand, patient individual data can be collected, compared and learned from, while on the other hand, many electronic health record systems are based on structured data, ontologies and defined vocabularies, which makes the data easy to model from a Big Data perspective. These records allow to group data from different patients, which in turn allows to make informed decisions about possible successful treatments based on previous observations made in patients who show similar disease characteristics. In addition, this data also allows to make predictions about possible future health outcomes (Kurt et al., 2008), which might allow patients and healthcare providers to instigate possible countermeasures (in time) to potentially prevent or minimize adverse events. Rajkomar et al. (Rajkomar et al., 2018) used a deep learning method based on electronic health data to predicted multiple endpoints of hospital admitted patients. Such endpoints included medical diagnosis, mortality, re-admission and length of stay. While diagnosis and mortality predictions can be used to improve patient care, re-admission and length can help in optimizing and pre-planning hospital procedures, such as staffing and bed distribution. Gu et al. (Gu et al., 2021) used patient data to predict the likelihood for a patient to not follow their prescribed medication schedule, based on data gathered from home devices which monitor medication injection. The efficacy of drugs and their safety can also be affected by individual differences. It is already widely acknowledged that drugs can have different effects based on the patient's gender (Gandhi et al., 2004; Sharifi et al., 2021), but also other genetic, ethnic or environmental differences can impact the suitability of a drug or drug combination (Ivanov et al., 2014; M. R. Nelson et al., 2016; Ramamoorthy et al., 2015). Therefore, considering these parameters during treatment decisions, and selecting treatment plans based on individual responsiveness to a drug, can lead to higher medication effectiveness, reduce the risk of adverse outcomes and as a result decreased healthcare cost overall (Fröhlich et al., 2018).

### 2.2.3 Hospital/ Healthcare Optimization

Process optimization through data analytics is a common practice across all industries. Hospitals, healthcare providers or nursing homes are at their core “businesses”, build on complex processes, which can be improved by identifying bottlenecks, workflows and optimizing implemented processes. This optimization can lead to better care, reduced costs and in general more relaxed staff. By analysing workflows and processes, the distribution of equipment can be optimized and adjusted, as well as potential machine (part) failure can be predicted ahead of time, which allows timely replacement and in turn improves safety, availability and performance (Aboul-Yazeed et al., 2017; Kovačević et al., 2020). In large cities patients can automatically be distributed across facilities based on bed, staff and experts available as well as within a facility a patient can be assigned the most optimal available bed based on the features of the room/ bed and the patients' requirements (Ceschia & Schaerf, 2011; Taramasco et al., 2019). Emergency calls can be automatically classified based on their severity and a suggestion of the best suitable response team, based on their knowledge, experience, closeness and availability can be made, which allows also non medically trained personal to take emergency calls (Ferri et al., 2021).

### 2.2.4 Image/ Disease Recognition

Diagnosis is often based on imaging data, being it radiographic produced images (X-ray), Magnetic resonance imaging (MRI), optical coherence tomography (OCT) or photographic images (Giarratano et al., 2020; Jaber et al., 2020; Phung et al., 2022; Salvatore et al., 2014; Z. Zhang & Sejdić, 2019). Computational analysis and classification of such images can make it possible to identify disease stages before a human expert would be able to make a diagnosis, reduce human error as well as allow to detect early changes, which may even occur in tissues/ areas unrelated to the disease phenotype (Giarratano et al., 2020; Haghanifar et al., 2020; Hou & Gao, 2021; M. Liu et al., 2015; Phung et al., 2022; Salvatore et al., 2014; Veena et al., 2017).

### 2.2.5 Clinical Trials

Clinical Trials are expensive and time-consuming endeavours. While they are necessary to ensure the success rate and safety of a drug, clinical trial design can be

optimized by analysing past trials and their set-up (A. Li & Bergan, 2020; Mayo et al., 2017). In addition, data from similar drugs or phenotypes, can be used to enrich or predict outcomes, which can minimize the amount of late-stage trial failures (Z. Chen et al., 2022). Of especial interest can be terminated clinical trials or trials that have been suspended before their intended end data. This data can help in learning from previous mistakes, not to put resources into trials, that may fail due to design, condition or compound as well as support the re-design/ improvement of trials that may be resumed later (Zame et al., 2020). Chen et al. (Z. Chen et al., 2022) created the Clinical Trials Knowledge Graph (CTKG), which is a large scale data collection of clinical trial information modelled as a Knowledge Graph and suggested its possible application for drug repositioning, to identify similar medical entities, such as phenotypes, drugs and studies, or for the designing of future clinical trials.

The COVID-19 pandemic has put clinical trials in front of new challenges, where a fast discovery of successful and safe drugs was needed. Virtual clinical trials combine patient data across hospitals, states or countries and group them based on treatment and patient descriptors into virtual treatment and control groups. While virtual clinical trials may not yet compete with classical clinical trials, they can suggest the first compounds to be tested in clinical trials and in result speed up the process and increase success rate (Zame et al., 2020). This information can be further enriched with data known about the compounds, phenotype or similar entities, to predict the most likely to be successful drugs. Observational patient data can also be leveraged to emulate randomized clinical trials, if these are not available (Admon et al., 2019; Franklin et al., 2021; Hernán & Robins, 2016).

## 2.2.6 In Toxicology, Drug Design & Chemical Safety

Toxicology, drug design and chemical safety are traditionally based on the principle of trial and error. Through testing it is evaluated what the compound does from a phenotypic point of view, understanding the molecular mechanisms behind the observed outcomes was not a priority. With the rise of –omics technologies, understanding instead of observing has become a research focus (Federico et al., 2020; Kinaret, Serra et al., 2020; Serra et al., 2020). When adding Big Data, data analytics and artificial intelligence into the mix the aim is to predict molecular mechanisms and phenotypic outcomes before any time and cost extensive experiments have been performed. In addition, it allows to suggest or even design compounds for a desired outcome, instead of relying on the classical method of trial and error.

Integrative approaches focus on combining computational and data driven power with experimental validation. They are built on the assumption that while the scientific community has already created enormous amounts of data, they are by far not enough to completely rely on computational predictions. Especially as such models become weaker the less prior knowledge around a data point is available. Therefore, they focus on using computational predictions to pre-filter potential compounds (Serra, Fratello, et al., 2022).

## Compound Design, Drug Repurposing and Drug Adverse Outcome Monitoring

*De novo* drug design focuses on going from a desired (phenotypic) outcome to the optimal compound, instead of going from a compound to its observed phenotypic outcome. While the aim of *de novo* drug design is to create novel compounds, during drug repurposing the aim is to identify already existing compounds, which could be used as treatment for additional phenotypic conditions (Mingyang Wang et al., 2022). Both methods rely on large scale data being available, covering data such as compound structural properties, target information, -omics data, associated molecular processes or textual based information.

The *de novo* drug design process is mostly based on structural representations of compounds and the suitability of the newly created compound is scored based on its binding affinity to the desired target. The structural information can be two-dimensional, three-dimensional, can be based on whole compounds, substructures or molecular descriptors and properties of the training compounds (Bai et al., 2021; Domenico et al., 2020; Mouchlis et al., 2021; X. Tong et al., 2021; Mingyang Wang et al., 2022). The use of Big Data libraries in combination with machine learning methodologies, allows the inclusion, screening and learning across millions of existing compounds to propose compounds with desired functions, such as specified protein and phenotypic targets, the minimization of adverse effects or specific physical properties for their industry application (Bai et al., 2021; Fang et al., 2023; Meyers et al., 2021; Mouchlis et al., 2021; Mingyang Wang et al., 2022).

Drug repurposing studies can be performed with a variety of methods and on a diverse set of data (Wieder & Adam, 2022), common data types are however protein protein interactions (PPI), drug targets, compound structural information, -omics and/ or clinical data (Afzaal et al., 2022; Cheng et al., 2019; J.-H. Gan et al., 2023; C. Xu et al., 2018; Xu Zhou et al., 2020). Zhou et al. used micro ribonucleic acid (RNA) drug associations (drug microRNA regulation (X. Liu et al., 2013) or



microRNAs impacting gene expression of gene products relevant for the drug function (Rukov et al., 2014) ) to create a drug – drug network, by determining if two drugs share statistical significantly microRNAs. This network is further enriched with drug – disease relationships and the set of therapeutic drugs for a disease is used as seed nodes for random walks (s. chapter 4). The most often visited drugs are considered as potential repositioning candidates for the selected disease. XU et al. developed the core-signature drug-to-gene software, which builds cell signatures for cancer cells based on occurred gene mutations and compares these with drug gene signatures computed from microarray data, retrieved from the Connectivity Map (CMap) dataset. Cheng et al. (Cheng et al., 2019), combined PPI, drug gene target, drug drug interactions, structural drug information, protein sequence information and gene expression profiles with additional information, such as Gene Ontology or phenotypic data to suggest drug combinations for specific diseases. DrugRep , is a virtual screening-based drug repositioning software, offering receptor and ligand-based screening, based on compound and protein structural information. Afzaal et al. (Afzaal et al., 2022) performed virtual screening on the ZINC database to identify compounds, which could act as possible human telomerase reverse transcriptase inhibitors, based on docking analysis of compounds structurally like compounds known to bind to human telomerase reverse transcriptase. Serra et al. combined multiple approaches and data types, such as gene co-expression network analysis , dose-response analysis , differential gene expression analysis (Federico et al., 2020) and QSAR modelling to identify possible repositioning candidates for the treatment of COVID-19.

Electronic health records provide a vast amount of data for drug repurposing studies, since they allow the testing of drug repurposing hypothesis across a large group of patients and time (Zong et al., 2022). Additionally, mining the large-scale information available, can help in detecting hidden associations between drugs and phenotypes or help in discovering complementary or interfering drug combinations (Y. Wu et al., 2019). The information can further be integrated with other data, such as structural/ chemical data points, experimental data, such as –omics data or biological knowledge, such as PPI, Adverse Outcome Pathways (AOP) (Ankley et al., 2010; Saarimäki, Fratello, et al., 2023), Gene Ontology (The Gene Ontology Consortium, 2021) terms or pathways (Goldstein et al., 2018; M. Zhou et al., 2021).

Not all drug adverse reactions or drug interactions can be identified during clinical trials. Therefore, continuous collection of data after its release is important. Social media and online platforms in addition to medical and patient reports allow the collection of large-scale data in combination with additional meta-data about

patients, such as lifestyle, which often may not have been reported in official reports. These data can be collected across nations and allows the scanning for potential adverse effects in specific sub-populations, the detection of drug drug interactions or the identification of outside influences on the efficacy and safety of a compound (Coloma et al., 2011; X. Wang et al., 2009). Adverse effects, in addition, can already be predicted before a compound is released to the market, allowing to identify possible harmful effects before large groups of patients can be affected. These methods often rely on large amounts of available drug, protein and phenotype data (Galeano et al., 2020; Liang et al., 2023; Nguyen et al., 2021; Pancino et al., 2022; H. Zhou et al., 2020). Galeano et al. (Galeano et al., 2020) developed a computation matrix decomposition-based model, able to predict side effect frequencies for unknown adverse drug outcomes from a small set of already known adverse effect frequencies. This method can be used in combination with GSEM (Galeano & Paccanaro, 2022), a matrix completion-based tool, which can predict possible side effects for a drug, given a small set of already identified adverse effects for a compound. DSGAT (X. Xu et al., 2022), uses the drug molecular graph to predict side effect frequencies, trained on known drug – side effect relationships. However, since DSGAT assumes that similar drugs will have similar adverse outcomes and vice versa, it can also be applied to drugs for which no adverse effects or their frequencies are reported yet.

## Mechanism of Action, Compound Toxicity and Toxicogenomics

With the development and use of different compounds in the industry and medicine, ensuring their safety for human exposure and the environment is essential. However, the sheer amount of existing compounds and ever newly developed compounds makes manual testing from a time and cost perspective impossible (Pavel, Saarimäki, et al., 2022; Serra et al., 2020). Therefore, it is of utmost importance to use alternative (computational) methods to predict the possible adverse effects of an exposure as well as to understand its mechanism of action (MOA). Commonly used data used for toxicity prediction and the characterization of the MOA are –omics data, such as gene expression, structural compound information or other biological data, such as Gene Ontology or pathway gene sets, AOPs and phenotypic or genetic data points (A. Liu et al., 2023; Serra et al., 2020).

QSAR and Structure Active Relationship (SAR) models use a compounds structure and known associations to predict the potential outcomes of structural similar compounds, where SAR models create qualitative relationships between a compounds sub-structural area and its possible toxicity, QSAR models create

functions describing a compounds structural descriptor in relationship to its physiochemical, toxicological as well as biological endpoints (Ruiz et al., 2012). Overall, the underlying assumption is that structural similar compounds behave similar to each other (Serra et al., 2020) under the constraint of the selected endpoint. Chen et al. (J. Chen et al., 2021) trained a graph convolutional neural network on the Tox21 (Richard et al., 2021) data, based on the simplified molecular input line entry system (SMILES) representation of compounds and known endpoints measuring seven nuclear receptor signals as well as five stress response indicators. In Sharma et al. (Sharma et al., 2023), molecular fingerprint signatures and SMILES embeddings are used to model *in vitro*, *in vivo* as well as clinical toxicity with multi and single task deep neuronal networks. Chirico et al. (Chirico et al., 2021) released QSARINS-Chem, a software to perform QSAR analysis, based on multiple different QSAR models, to predict different endpoints, such as toxicity, physio-chemical properties or environmental persistence.

AOPs model chains of events, so called Key Events (KE), in a consecutive order, leading from a molecular initiating event (caused due to an exposure) to an adverse outcome in an organism or population (Ankley et al., 2010). AOPs are a relatively new model and only a few hundred manually created AOPs are currently available. However, through being modelled as a network, it is possible to infer outcomes or travel backwards from an observed outcome by identifying a matching event block in the chain (Kinaret et al., 2020). Bell et al. (Bell et al., 2016) developed a computational model, which based on toxicogenomic data creates computational predicted AOP modules. Wu et al. (Q. Wu et al., 2021) used the AOP framework to identify biological processes linked to drugs, through linking drugs to existing Key Events in the AOP network. Ball et al. (Ball et al., 2021) combined Key Event information with biological assays and chemical structural information to make predictions about adverse outcomes.

### **Toxicogenomics**

The field of toxicology focuses on understanding how different compounds can affect humans and the environment (Kinaret, Serra et al., 2020). Toxicogenomic approaches focus on studying molecular alterations with respect to a compound exposure, which can provide insights into the underlying molecular processes taking place, which helps to characterize a compounds MOA instead of observing the exposure effect on organism level, as done in traditional toxicology (Kinaret, Serra et al., 2020; Z. Liu et al., 2019). The most popular methods to measure gene expression are Deoxyribonucleic acid (DNA)-Microarrays, Polymerase Chain Reaction (PCR) and RNA-Sequencing, which both count the occurrences

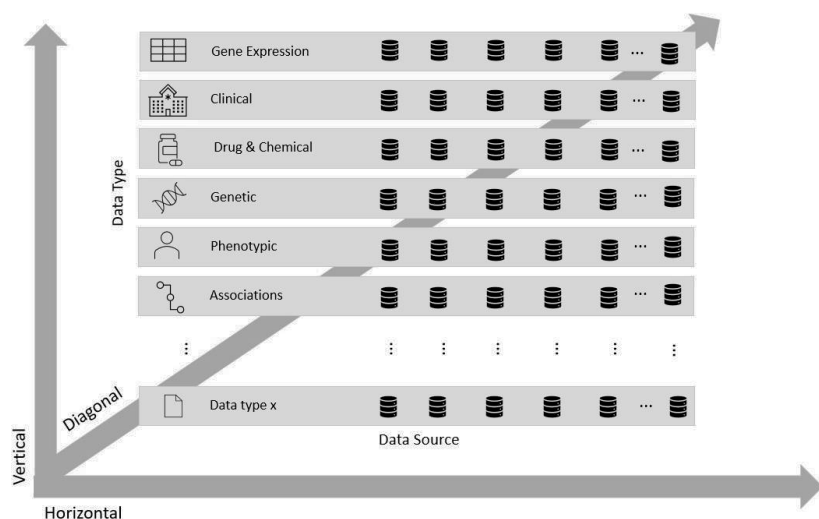
(expression) of specified genes (Kinaret, Serra et al., 2020; Z. Liu et al., 2019; Rao et al., 2018).

Such transcriptomics assays can be used to characterize a compounds MOA by analysing its induced transcriptomic alterations (Federico et al., 2020). Gardiner et al. (Gardiner et al., 2020) tested multiple classifiers, such as linear regression, K-nearest-neighbour and random forest to predict kidney disfunction on transcriptomics profiles from the Library of Integrated Network-based Cellular Signature (LINCS) 1000 dataset (Subramanian et al., 2017). Kohonen et al. (Kohonen et al., 2017) developed the predictive toxicogenomic space, which models cytotoxic effects based on transcriptomic profiles of the connectivity map dataset (Lamb et al., 2006). Fortino et al. (Fortino et al., 2022) analysed a set of 31 engineered nanomaterials (ENM), based on their induced toxicity levels, to identify physiochemical properties as well as molecular features that can distinguish between the toxicity classes.

### 3 DATA INTEGRATION

Big Data has large potential in the life sciences of which some cases have been outlined in the previous chapter. Many computational models exist and their predictive as well as analytical power for different scenarios have been shown. Many of the current models are trained and created on only a few data types and data sources while the data has been collected for a specific application scenario only (Pavel, Saarimäki, et al., 2022). However biological processes and events are of high complexity and in order to understand them in detail, many different data points, measuring and/ or describing a variety of (molecular) process or entities, need to be considered. Capturing all these with a single data type (or limited number of data types) is not possible. Therefore, combining large sets of different data types and sources can increase the predictive and modelling power as well as improve the understanding of underlying complex molecular processes, which not only focus on observable endpoints, such as dead & alive, but rather to understand why these endpoints appear. However, the data in the life sciences is by tradition highly fractured, lacking defined standards and methodologies across sub-fields (Leonelli, 2019; Pavel, Saarimäki, et al., 2022). Therefore, before the full potential of the data can be explored, data integration and modelling strategies, adapted to the complexity and diversity of life science data, need to be explored.

Data Integration (DI) describes the process of combining data coming from different sources or types into a single view or system. DI can focus on the integration of the same data type, different data types or both combined (Figure 3). DI is important to create large data sets, which combine data from multiple source systems, which can provide a more complete view of the problem under investigation than a single data source can.



**Figure 3** Vertical data integration, combines data across multiple data types, horizontal data integration combines data of the same type across multiple data sources and diagonal data integration combines data of different types across different data sources.

### 3.1 Horizontal Data Integration

Horizontal Data Integration describes the process of combining data of the same type, coming from different sources and possible in different formats (Argelaguet et al., 2021; Mihaylov, Nisheva-Pavlova, et al., 2019; Urbanski et al., 2019). So can for example the same entity be identified by different identifiers, the data can be created with different technologies or be of different quality (Pavel, Saarimäki, et al., 2022). Horizontal data integration can improve the robustness of the data type, due to having multiple data sources, and possible technologies combined. This assumes that the more support a specific data point receives, the more likely it is to be true. In addition, horizontal DI can lead to the creation of a more complete picture of the data under investigation, due to possible covering a larger data space than a single data source would (Pavel, Saarimäki, et al., 2022). A similar concept is applied in the wisdom of crowds' principle (Marbach et al., 2012). Horizontal data integration is popular in patient studies, where data from multiple patients are combined to study a common phenotype (Ravera et al., 2021), to create a robust PPI network by combining data from multiple sources (Martha et al., 2011) or in the study of multiple transcriptomic datasets (Oestreich et al., 2022).

## 3.2 Vertical Data Integration

Vertical Data Integration describes the process of combining data of different types to gain a more comprehensive view of the entity (entities) under investigation (Argelaguet et al., 2021; Mihaylov, Nisheva-Pavlova, et al., 2019; Oestreich et al., 2022; Urbanski et al., 2019). For example, can multiple data about a phenotype, such as patient lifestyle, treatment and clinical data together with transcriptomics be combined to understand the phenotype under investigation in more detail as well as to form a more informed picture, than would be possible from single data points (Mihaylov, Kańdula, et al., 2019; Serra et al., 2019; M. Wu et al., 2021). Vertical data integration is popular in –omics studies, where the results of different omics experiments are combined (Serra et al., 2015; Ulfenborg, 2019).

## 3.3 Diagonal Data Integration

Diagonal Data Integration refers here to data integration that combines both aspects of horizontal and vertical DI (Argelaguet et al., 2021; Y. Xu & McCord, 2022). In result, it both supports robustness of individual data points and data types, as well as allows to form a comprehensive multidimensional view over data points. This is important, especially when Big Data models are used in computational approaches for chemical safety assessment, where a diverse set of data is needed to approximate organism level (Pavel, Saarimäki, et al., 2022). Diagonal DI is on the rise but due to its complexity not as widespread in the life sciences as horizontal and vertical DI but has become popular in the single cell sequencing field (Y. Xu & McCord, 2022).

## 3.4 Challenges of Big Data & Data Integration in the Life Sciences

Large-scale data integration across multiple data types and data sources poses some unique challenges in the life sciences which are not as strongly observable in the IT-industry.

### 3.4.1 Entity Identification and Mapping

By tradition the life science research field is highly fractured (Leonelli, 2014, 2019; Marx, 2013) with multiple different sub-disciplines. As a result, there are multiple

standards, with respect to methodologies, definitions and naming systems available. Even within the same sub-field differences based on research language or location can be observed. An example for this is the difference in gene naming and gene definition between the United States of American Entrez system (Maglott et al., 2011) and the European Ensembl system (Cunningham et al., 2022), for which no 1-1 mapping exists. However, solving this identification challenge is not a technical but a semantic task, where common terminologies, such as ontologies, need to be defined, accepted and enforced across research fields, institutes and countries (Pavel, Saarimäki, et al., 2022). The current standard of using whatever identifiers the researcher in question is used to, as well as a preference for using language-based terminology, instead of fixed identifiers, makes the large-scale data integration across datasets, and especially across disciplines highly challenging. While using gene symbols or phenotype names are easier to understand by human readers, they are not strictly defined. Not for all genes official gene symbols exist, and a gene can refer to for example multiple proteins, as well as can variations in the spelling be observed. While for a human reader small spelling differences are acceptable to a computer these are different strings (Locke et al., 2021). Natural Language Processing (NLP) can identify these matches with a confidence score, however it is easily thrown by the highly similar naming used for different genes, where a single character difference can either indicate that it is the same entity or may also be a different entity. Such conflicts cannot be resolved without human interaction and correction as well as an acceptance of mistakes, based on relying on confidence scores during the entity mapping process.

Entity identification and mapping is further complicated by terminologies that are language dependent, such as phenotypes (diseases), where especially for clinical data, the data is reported based on the terminology and language common to the location the data is produced in. Further not all phenotypes are clearly defined or are reported in different granularities across organisations.

NLP based entity recognition is further complicated by the lack of large, annotated training data, which can lead to poorly generalizable models and therefore introduce a data bias by identifying specific terms more than others, due to their availability in the training data (Leaman et al., 2015; Locke et al., 2021).

The entity identification and mapping challenge is one of the fundamental causes why large scale inter-disciplinary data integration in the life sciences is still missing and many of the current projects are built on a small number of source systems only,



due to the manual effort needed in supervising or correcting the automatic entity identification (Pavel, Saarimäki, et al., 2022).

### 3.4.2 Data Design does not take Big Data Applications into Account

Big Data analytics are based on the processing of large amounts of data that cannot be performed manually. This implies that data would need to be created in machine readable formats with the goal of data integration, re-use and computational processing in mind. However, for many researchers and life science sub-fields, such as clinicians, this is not the focus of their work (Pavel, Saarimäki, et al., 2022). Further the heterogeneous landscape of information and recording systems used across health care providers and nations creates information silos that are not directly interoperable between each other (Queralt-Rosinach et al., 2022). However, in recent years there have been efforts to unify terminology used across systems by defining ontologies to be used by all participants (Queralt-Rosinach et al., 2022).

In addition, many research outputs are often not published with computational readability or in general re-use in mind, making the output nearly impossible to use in Big Data analytics studies. While the emerging FAIR (Findable, Accessible, Interoperable and Reproducible) principles are a step in the right direction, to enforce the documentation and reporting of digital data (Martínez-García et al., 2023), the focus is on individual dataset and not on reporting data in such a way that it is re-useable in large scale data integration projects (Pavel, Saarimäki, et al., 2022). This means metadata can be reported differently for the same data types (Hughes et al., 2023) or data points measured can vary as well as is there no standard on what “good quality” data is (Saarimäki et al., 2022). Hughes et al. (Hughes et al., 2023) further criticised the lack of incentive for researchers to publish their digital data with complete and re-useable metadata, which significantly limits the findability and reusability of data as well as leads to duplicate efforts.

Lastly the lack of taking computational use into account leads to a lack of reporting of negative data points (Pavel, Saarimäki, et al., 2022). Supervised machine learning tasks are relying on the availability of positive and negative samples, however in the research community there is a focus on only considering positive results as noteworthy (Maloney et al., 2023; Nimpf & Keays, 2020), while from a machine learning point of view both are valuable and their lack can lead to data driven biases in the analyses and trained models.

These topics make large scale data integration and Big Data analytics highly challenging in the life sciences. However, especially in the chemical safety and drug development process, alternative methods to large scale experimental approaches are highly sought after, as a result from a time, cost and ethical point of view.

## 4 NETWORK ANALYSIS IN THE LIFE SCIENCES

Network models and their analysis have become popular in many life science sub-fields, due to the fact that many data types are by nature reported as networks, such as PPI (Steven Wang et al., 2022), regulation (Pratapa et al., 2020) and metabolic networks (Christensen & Nielsen, 2000), as well as pathways (Jassal et al., 2020; Kanehisa et al., 2017) and ontologies (Köhler et al., 2021; The Gene Ontology Consortium, 2021). In addition, associations and correlations are often only different representations of a network data structure (Marwah et al., 2018; Pavel, Saarimäki, et al., 2022). In addition, the modelling of data as a graph makes it possible to analyse data points with respect to their surroundings and the whole system, instead of as individual entities. Network analyses have found especially application in the analysis of protein interactions (Jeong et al., 2016; N Przulj et al., 2006; Simões et al., 2012; Steven Wang et al., 2022), gene co-expression networks (Marwah et al., 2018; Odibat & Reddy, 2012; Song et al., 2019; Yuan et al., 2017) or for drug repositioning (Badkas et al., 2021; Federico, Fratello, et al., 2022; Zeng et al., 2019; Xu Zhou et al., 2020).

### 4.1 Network Inference from Transcriptomics

Gene expression data coming from transcriptomic assays can not only be used to study individual genes, but also to investigate the interactions between them as well as to model the whole system (Federico, Pavel, et al., 2022; Pavel, Serra, et al., 2022). Gene regulatory networks model regulators and their gene targets as nodes and edges represent their relationships (Y. Gan et al., 2022). In gene co-expression networks genes are modelled as nodes and edges indicate if a pair has a (significant) co-expression (i.e., shows similar expression patterns across samples and/ or conditions) relationship (Marwah et al., 2018; Pavel, Serra, et al., 2022). Unsupervised network inference algorithms mostly rely on principles from information theory, where the correlation and/ or mutual information is measured between the expression profiles of two genes and a relationship is inferred if a significant correlation between two genes exists (Pavel, Serra, et al., 2022; Zhao et al., 2022). Multiple different algorithms exist which try to estimate significant edges

in a more sophisticated manner, than setting a significance threshold (Butte & Kohane, 2000; Faith et al., 2007; Langfelder & Horvath, 2008; Margolin et al., 2006; Marwah et al., 2018; Meyer et al., 2007; Zhao et al., 2022). However, correlations are by nature bi-directional, which does not allow to infer a regulation direction when computing a regulatory network. Therefore, supervised approaches have been proposed, such as DGRNS (Zhao et al., 2022), which adds time information onto the gene expression data or GRADIS (Razaghi-Moghadam & Nikoloski, 2020), which uses prior knowledge of transcription factors.

Gene co-expressions can be used to study the system response to a specific condition or can be compared across conditions, tissues, cell types or organisms to identify similarities and changes in gene gene relationships (Anglani et al., 2014; Federico, Pavel, et al., 2022; P. Kinaret et al., 2017; Y. Liu et al., 2019; Odibat & Reddy, 2012; Ovens et al., 2021; Pavel, Serra, et al., 2022; Pierson et al., 2015; Song et al., 2019; Y. Yang et al., 2014; Yuan et al., 2017). These insights can for example be used to characterize and/or compare a compounds MOA (Koenig et al., 2021; W. Liu et al., 2018), identify possible drug targets or key players in a condition under investigation (Federico, Pavel, et al., 2022; Hasankhani et al., 2021; Y. Liu et al., 2019; W. Li et al., 2020; Song et al., 2019; Tanvir & Mondal, 2019; Yuan et al., 2017), or to outline cell, tissue and organism differences (Eidsaa et al., 2017; Ovens et al., 2021; Pierson et al., 2015; Y. Yang et al., 2014).

The different biological networks are often analysed and/ or compared based on different network metrics and methods.

## 4.2 Network Metrics and Their Applications in Toxicology and Pharmacology

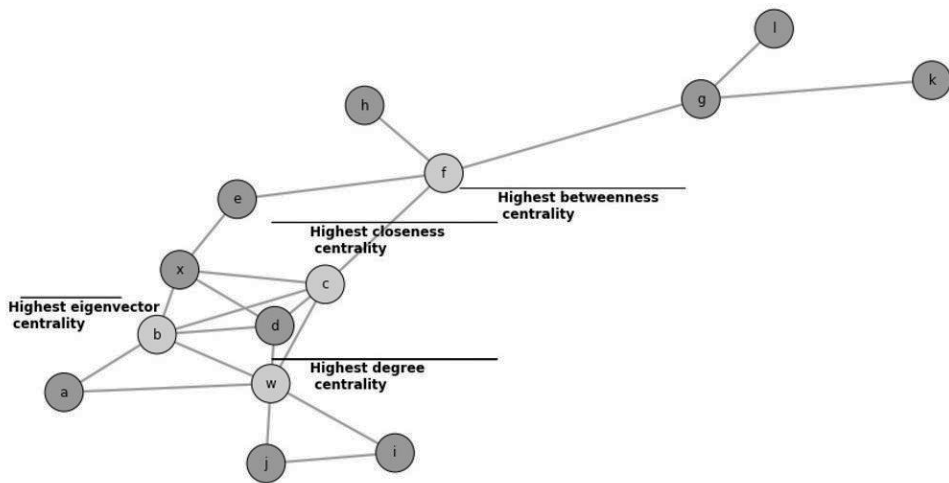
Topological metrics make use of the network structure to score nodes and edges based on their importance in the network (Pavel, Serra, et al., 2022). These measures either take local network properties into account or the whole network. Scoring the importance of a node in a network is a popular method applied onto biological networks. Degree centrality evaluates the importance of a node by the number of edges it contains, while closeness centrality scores the node based on how close it is to any other node in the network. Eigenvector centrality measures the influence of a node based on its connection to “influential” nodes and betweenness centrality evaluates the importance of the node based on its contribution to the information flow in the network (Pavel, Serra, et al., 2022)

(Figure 4). Shortest paths are the least expensive (based on steps or edge weights) path that can be taken between two nodes in the network (Ren et al., 2018; Simões et al., 2012), while random walks are random steps taken in the network from a starting node (Lao et al., 2011; Newman, 2005; Rosvall & Bergstrom, 2008). Cycles are loop structures in the network (Giarratano et al., 2020; Paton, 1969) and graphlets are small sub-graphs that can be used to describe the topology of a network (Hayes et al., 2013; Przulj, 2007; Sonmez & Can, 2017). Examples of network metrics and their possible application on biological networks are listed in Table 1.

Cheng et al. (Cheng et al., 2018) performed drug repositioning on a human PPI network, by estimating the shortest path distance between a drug target and a disease protein and estimated their significance by comparing their shortest path distance to the distribution of shortest paths in the network. A similar approach is applied by Zhou et al. (Y. Zhou et al., 2020) where drug repositioning for COVID-19 is performed on a PPI network, by estimating the shortest paths between virus-host interactors and drug targets on the PPI network. Manczinger et al. (Manczinger et al., 2018) performed drug repositioning by modelling the impact of a drug via its target on a PPI network and evaluating its efficacy based on downstream affected proteins. Guo et al. (Guo et al., 2022) identified genes associated with the studied phenotype (unstable carotid atherosclerotic plaques) by computing differentially expressed genes across multiple transcriptomic datasets and using them to construct a PPI network via STRING (Szklarczyk et al., 2019). The most central genes are identified based on multiple centrality metrics and miRNA, transcription factors as well as drugs targeting the hub genes are identified and used to identify the gene targeted by most miRNAs/ transcription factors and drugs. A similar approach of identifying genes associated to a studied condition, based on the central proteins in a PPI network, constructed from condition related differential expressed genes has been used to study for example heart failure (Tu et al., 2022), diabetes (Prashanth et al., 2021), and breast cancer (Tang et al., 2019). Sonmez et al. (Sonmez & Can, 2017) grouped tissue specific PPI networks, by describing each network based on their graphlet counts.

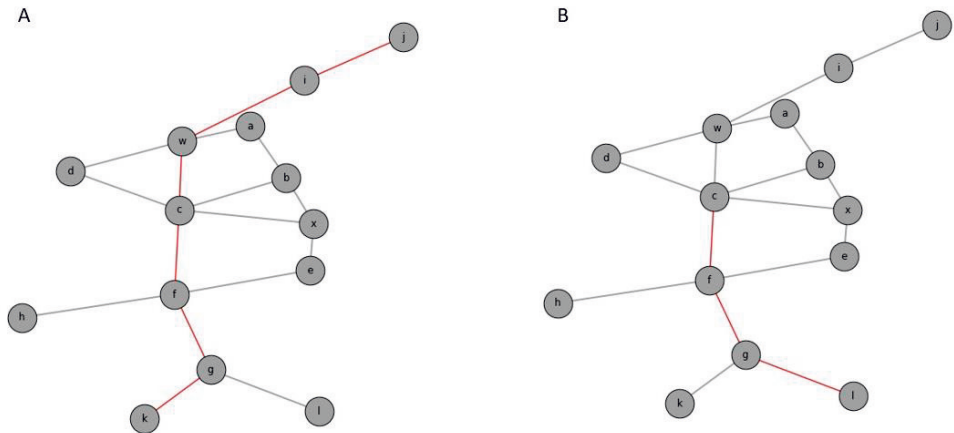
**Table 1** Examples of network metrics that can be applied to biological networks.

<b>Metric</b>	<b>Category</b>	<b>Explanation</b>
Degree	Connectivity of a Node	Number of connections of a node, for directed networks it can be measured as in-degree and out-degree, i.e., incoming and outgoing edges. This metric can identify the hub nodes in a network. In cancer therapy they may be considered good targets to destroy the whole system (Federico, Fratello, et al., 2022).
Closeness Centrality	Node position in the network	Measures the distances from a node to all the other nodes in the network. High centrality nodes in a set of disease related genes, may be suitable drug target candidates (Pavel, Serra, et al., 2022).
Betweenness Centrality	Node position in the network	Measures how important a node is for the information flow in a network. It measures how often a node is on the shortest path between any two nodes in the network. These genes may be good targets when genes with the most influence between two sets of genes are sought (Federico, Pavel, et al., 2022; Pavel, Serra, et al., 2022).
Eigenvector Centrality	Node position in the network	Measures a nodes importance based on its connectivity to other important nodes in the network.
Shortest Path	Path in the network	The minimum resource path to take in the network between two nodes (Pavel, Serra, et al., 2022).
Random Walk	Path in the network	Random walks are walks performed on the network, where each step is performed at random, edge weights can be considered, corresponding to the probability a step is taken (Lao et al., 2011; Newman, 2005; Rosvall & Bergstrom, 2008).
Cycles	Loop in the network	Cycles in the network can have different meanings based on the type of network. In regulation network cycles can for example indicate feedback loops or can be used to describe the topological structure of a vascular network (Giarratano et al., 2020; Paton, 1969).



**Figure 4** Node w has the highest degree centrality due to being the node with the most edges, b the highest eigenvector centrality due to being connected to high value nodes, such as w and d, f the highest betweenness centrality due to being the only path between g, i, k and all the other nodes and c the highest closeness centrality due to being closely connected to many of the nodes in the network. Figure adapted from (Pavel, Serra, et al., 2022).

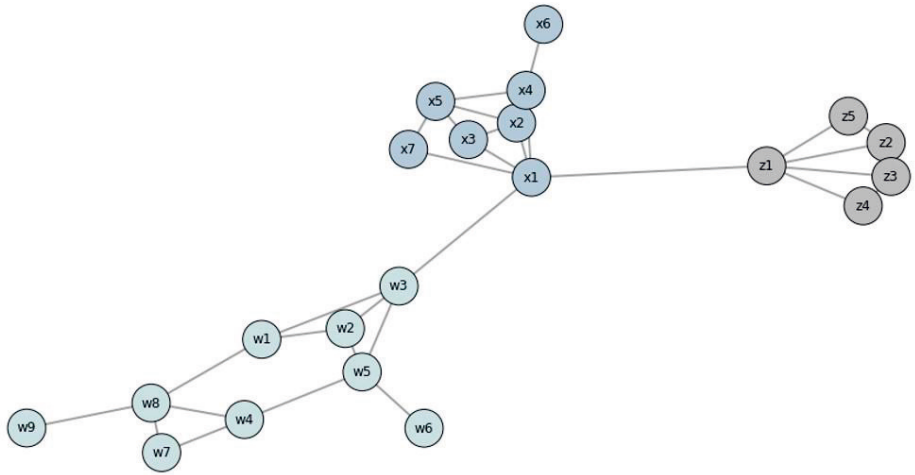
Global network metrics describe the graph structure and its connectivity. Such metrics are for example the graph radius, which is the smallest of all the longest shortest path of each node in the network, the diameter, which is the longest shortest path in the network (Figure 5), the clustering coefficient, which measures how connected the graph structure is based on the tendency of triplet structures forming triangles or the graph density, which is the fraction of existing edges out of all possible edges (Pavel, Serra, et al., 2022). In addition, local metrics can be calculated for each node and their distribution across the whole network can be used to describe the global topology of the network. Graphlet distribution and random walk-based path distribution, as well as the loop structure of the network are used by Giarratano et al. (Giarratano et al., 2020) to create retinal biomarkers applied to the vascular network of retinal images (OCTA images) to identify patients with diabetic retinopathy and chronic kidney disease.



**Figure 5** A) graph diameter (red) and B) graph radius (red).

Often not only single nodes are of interested but groups of nodes. In gene co-expression networks identifying communities can help in detecting active biological processes (W. Li et al., 2020; Marwah et al., 2018; Tanvir & Mondal, 2019). A network community (Figure 6) is described as a densely connected subgraph, i.e., a group of nodes that interact with each other more often than with other nodes in the network (Linhares et al., 2020). Kinaret et al. (P. Kinaret et al., 2017) showed that using communities on gene co-expression networks allows to approximate the MOA (based on gene set enrichment) between *in vitro* and *in vivo* engineered nanomaterial exposures. Combining community detection on gene co-expression networks with computing community hub genes to identify disease related genes and pathways, based on pathway enrichment, has been applied to study multiple diseases, such as cancer (Yuan et al., 2017), bipolar disorder (Y. Liu et al., 2019) and atrial fibrillation (W. Li et al., 2020).





**Figure 6**  
al., 2022).

A network with three communities, w, x and z. Figure adapted from (Pavel, Serra, et

## 5 KNOWLEDGE GRAPHS

Knowledge graphs are reasoning engines, which are applied on top of a graph data model (Figure 1), which have recently gained attention as possible data models and prediction engines for Big Data analytics in the life sciences (Pavel, Saarimäki, et al., 2022).

### 5.1 Advantages of a Graph Data Model

Graph database management systems, as well as other No-SQL (non-tabular) database management systems are schema-free. This means that the data model does not need to be defined in advance, and therefore can evolve with the data and its applications (Pavel, Saarimäki, et al., 2022) and in addition, it allows for gaps in the data. However, the schema free nature implies that more responsibility is put on the database administrator to ensure integrity of the data (Pavel, Saarimäki, et al., 2022). Life science data is by nature prone to gaps, differences and is constantly evolving. Different data sets will report different variables, different technologies measure and evaluate different parameters and between sub-disciplines even the same term may not refer to the same entity or standard. Therefore, a flexible and schema free data model is needed when creating an integrated and expandable Big Data knowledge base for the life sciences.

In contrast to other data models, network data models put high emphasis on connections and paths instead of entities. This is in accordance with data types in the life sciences and their research questions, where often relationships, associations or correlations between entities are of high interest, instead of the entities themselves. Especially the field of systems biology aims at understanding molecular processes, their interactions and impacts through the modelling of networks (Albert, 2007; Arrell & Terzic, 2010; Serra et al., 2019; Yan et al., 2018). But not only data that is commonly analysed or represented as networks is of advantage to be modelled in a network-based data model. Any correlation, association, data represented in a tabular format or as a matrix can be seen as a representation of a network and therefore easily be translated into a network data model.

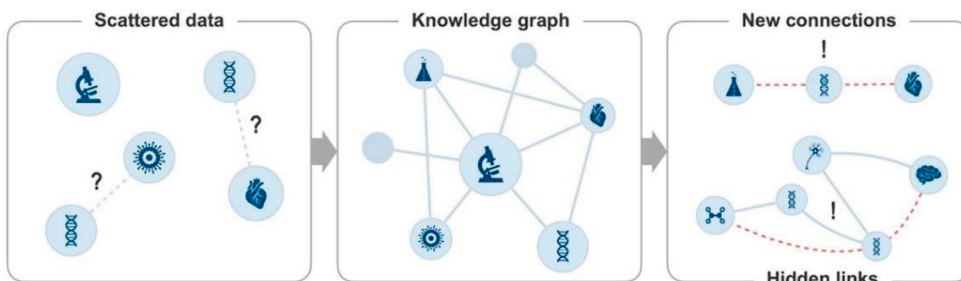
In addition, graph representation of the data allows to analyse the data in addition with network-based metrics (chapter 4).

### 5.1.1 Leveraging the Knowledge Graph Topology

The individual data layers, combined data layers, subnetworks or the whole KG network structure can be used to gain insight into the data modelled in the KG. Centrality metrics (chapter 4) allow to identify the entities in the network most connected, which can indicate high relevance or importance with respect to the modelled data but also can suggest knowledge biases existing in the available data. For example, cancer is a highly studied disease which results in more data available, while a rarer disease has less data available and as a result will have weaker connected entities. However, in a phenotype centred KG this would not mean that these diseases are for example associated with less genes, chemical exposures or mechanism of actions but only that there is less knowledge available. Community detection on a KG can again be performed on a subgraph of the KG or the whole and can for example be used to identify similar entities or entities that have a strong correlation. In a clinical KG communities can be used to identify patients that are highly similar with respect to their diagnosis, treatment or epigenetics (Z. Chen et al., 2022). Path based metrics, such as shortest paths and random walks are often directly applied during querying to retrieve direct and indirect information about the query entity.

#### Hidden Links

Hidden links represent information that is contained across multiple data sources and points but only becomes visible when these data are integrated into a common system (Pavel, Saarimäki, et al., 2022). The network-based structure of a KG makes the identification of these hidden links visible due to showcasing them as paths in the network as shown in Figure 7, which can be queried directly by the user instead of performing complex joins across multiple tables as would be needed in traditional relational database models.



**Figure 7** Shows the integration of data into a KG and how hidden links can become visible in a network data model. The figure is taken from (Pavel, Saarimäki, et al., 2022).

## Assessing Data Point Quality via Network Topology

It is commonly accepted that with Big Data comes the risk of adding erroneous data points, however the general idea is that individual false data points are outweighed by correct ones (Pavel, Saarimäki, et al., 2022). In the life sciences the concept of true and false is less strongly defined, since for many insights numerous variables affect the results, ranging from the experimental setup to individual patient variability. The topology of the KG can help in assessing the quality and likelihood of individual entities and relationships. The underlying assumption is that similar entities should be close in the network, and therefore relationships can be scored based on how many similar entities they are connected to, which allows the identification of possible outliers, while entities can also be scored on their connectivity profiles as well as node and edge labels can be assessed or suggested based on the local surrounding information (Lao et al., 2011; Pavel, Saarimäki, et al., 2022; Steenwinckel et al., 2022).

## 5.2 Knowledge Graphs as Predictive Engines

Data analytics and Big Data are often used as basis to make predictions about the world under investigation or future behaviours. A graph data model allows to make use of edge prediction algorithms as well as classical machine learning and prediction algorithms. Triangle closure or common neighbours is one of the simplest edge prediction algorithms. It assumes that if a node has a connection to two other nodes it is likely that these two other nodes also have a connection (Koutrouli et al., 2020; Leskovec et al., 2008). Through node/ graph embedding the graph space can be translated into a vector space (Grover & Leskovec, 2016; Pancino et al., 2022; Pavel, Saarimäki, et al., 2022; Steenwinckel et al., 2022). This allows to use classical vector-

based machine learning approaches, such as logistic regression (Serra et al., 2020). One of the most common node embedding algorithms is Node2Vec (Grover & Leskovec, 2016), which is based on Word2Vec (Mikolov et al., 2013) and its Skip-Gram model. The Skip-Gram model is used to estimate the probability of a word given another word and is trained over a specified window size on sentences. In Node2Vec sentences are replaced by random walks and the algorithm translates each node into a vector of dimension  $x$ , where nodes that appear more often on the same random walks are placed closer in space than nodes that do not appear on the same random walks (Grover & Leskovec, 2016).

### 5.3 Examples of Knowledge Graphs in Toxicology, Chemical Safety and Drug Development

KGs have been used in different stages of the drug/ chemical development and safety assessment process and shown their potential for a diverse set of tasks across various sets of data types (Al-Saleem et al., 2021; Che et al., 2021; Z. Gao et al., 2022; Mohamed et al., 2020, 2019; Nováček & Mohamed, 2020; Ratajczak et al., 2022; Meng Wang et al., 2021; Zhankun Xiong et al., 2022; F. Zhang et al., 2021; R. Zhang et al., 2021; Y. Zhu et al., 2020). Many of these models are however based on single data sets or limited data types, due to the complexity and diversity of available data (chapter 3). KGs offer a framework that allows to investigate relationships between entities across a multitude of data layers, hence improving the insight and predictive power, by taking multiple views into account and reducing (possibly) existing data biases (Pavel, Saarimäki, et al., 2022).

Drug drug interactions can have severe impact on a patient and are often only noticeable after a drug has been released on the market and taken by large numbers of patients. Predicting these interactions (positive or negative) *in silico* can be used to reduce potential harm to patients as well as to identify new drug combinations for successful treatments (Karim et al., 2019). Abdelaziz et al. (Abdelaziz et al., 2017) developed Tiresias, a framework built on the construction of a KG on which drug similarities are computed across different data types and the whole KG structure to predict drug drug interactions via logistic regression. MUFFIN (Y. Chen et al., 2021) combines drug structural information with bio-medical data, in form of a KG, to predict drug drug interactions, by converting both the KG and the graph structural representation of the drug into a vector space, which combined, is used in a classifier model. Karim et al. (Karim et al., 2019) constructed a KG from different drug

related databases, such as DrugBank (Wishart et al., 2018) and PharmGKB (Whirl-Carrillo et al., 2021) on which graph embedding is performed to transform the graph into a vector space which is used as input for a classification model. The approach of creating a KG, embed it and using a classifier model on the embedding vectors to perform a link prediction task has also been applied by Wang et al. (Meng Wang et al., 2021) to predict drug drug interactions but can be applied to many different tasks (Z. Chen et al., 2022; Z. Gao et al., 2022; F. Zhang et al., 2021) and currently is one of the preferred methodologies to learn from KGs in the life sciences (Pavel, Saarimäki, et al., 2022).

The aim of drug repositioning is to re-use already existing and approved compounds for other applications. KGs have a large potential to integrate relevant data, ranging from drug structural information, protein targets, protein interactions, molecular processes to phenotypes on which drug repositing can be performed (Al-Saleem et al., 2021; Zhankun Xiong et al., 2022; Zeng et al., 2019). In a KG this translates to a link prediction task, which can, for a specific relationship type, be translated into a binary classification task of relationship exists or does not exist. Especially for the prediction of COVID-19 drug repositioning candidates have KGs become popular (Al-Saleem et al., 2021; R. Zhang et al., 2021) due to their flexibility with respect to data diversity and availability as well as the computational models that can be added on top of them. Che et al. (Che et al., 2021) used available information about COVID-19 to add to their medical KG on which drug disease interactions are predicted via a graph convolutional network model. Ge et al. (Ge et al., 2021) constructed a virus centred KG, comprising drug target, virus drug, virus protein, protein interaction, drug similarity and protein similarity information on which drug candidates are extracted based on a graph convolution model. This information is further enriched with literature-based information, drug gene expression perturbation profiles, clinical trial data and molecular docking. KG-Predict is a drug repositioning framework which comprises KG construction, KG embedding and the scoring of relationships for a specific question (Vashishth et al., 2020). GraphPK (Zhankun Xiong et al., 2022) combines drug disease associations, a drug KG and drug structural information, which are all transformed into a vector representation and fed into a multimodal neuronal network for the prediction of drug phenotype pairs. Wang et al. (Shudong Wang et al., 2022) developed KG-DTI, Mohamed et al. (Mohamed et al., 2020) used an embedding based method and Thafar et al. (Thafar et al., 2020) developed DTiGEMS+, which combines node embedding of the KG via node2vec (Grover & Leskovec, 2016) with classifiers to predict drug target interactions from which possible drug repositioning candidates can be inferred.

As discussed in section 2.2.5, large amounts of data are generated during clinical trials, which can be pooled and analysed to produce new insights into drug combinations, drug interactions, drug repositioning candidates and drug adverse outcomes as well as can be used to perform virtual clinical trials (Zame et al., 2020). The Clinical Trials KG (Z. Chen et al., 2022) is a KG constructed based on clinical trial metadata and results, which can be used as a framework to design clinical trials, perform drug repositioning, based on the concept of similarity between the drug and phenotype entity in the KG, or identify similar entities, such as studies or phenotypes. All information retrieval on the Clinical Trial KG is based on its node embedding and estimating the cosine similarity between entities to score their “relatedness”.

Chandak et al. (Chandak et al., 2023) developed PrimeKG, a KG that integrates 20 data-sources, covering diseases and relevant relationships, such as pathways, biological processes and protein perturbations. PrimeKG is focused on disease subtyping and disease clustering, which can be used to identify the most suitable possible drug for a patient's disease-subtype, by leveraging known information in the knowledge graph about closely related phenotypes.

Toxicity of chemical compounds is not only a concern when considering human exposure, but they are a risk for the environment and every living organism in it. KGs have been used to understand a compound's influence on its environment (Myklebust et al., n.d.) as well as to have all safety information about a compound readily available when needed in an integrated knowledge base (Zheng et al., 2021).

These wide range of possibilities to analyse and learn from the data makes a graph-based data model highly flexible, in addition to allowing analysis and prediction methods that are unique to a network-based data model. Further it allows to translate the whole network into a vector space, which makes it analysable with a wide range of methods coming from different research areas. In addition, the model makes it easy to generate sub-graphs which can be used to investigate specific problems and research questions, while the common data-model makes it re-usable and adjustable to other research problems, areas and hypothesis developments. However, the unique computational and data integration challenges have so far hindered the large-scale development of KGs in the life sciences and mostly resulted in small scale KGs only including data tailored to specific analysis question (Pavel, Saarimäki, et al., 2022).

## 6 AIMS OF THE STUDY

Big Data and data analytics are being increasingly used in the life sciences, especially in the medical sub-fields. However, due to different challenges, such as non-standardization and non-computational data representations, existing studies have been limited to specific use-cases and are based on limited data sources and data types. In my PhD, I investigated the possibility to integrate multidisciplinary data of different quality, standards and completeness by combining manual and automatic data curation processes. Here I present a highly flexible data model which can evolve with the data as well as showcases the usage of this integrated knowledge base across different use-cases. To showcase the applicability of the developed data model, different case studies from the field of toxicogenomic and drug repositioning are performed. However, the data model can be used for many different application scenarios and can be expanded or adjusted easily to other research domains if needed.

- 1. Creation of a unified Knowledge Graph for drug and chemical safety.**
  - a. Proof of Concept that life science KGs with billions of data points can be created across a multitude of data types, data sources and data standards for its application in chemical safety and drug development.
  
- 2. Development of computational methods adapted to computational aided toxicology research.**
  - a. Metrics that can be used to analyse and compare biological networks as well as to be applied on the data layers of the KG or in combination with data extracted from the KG.
  
- 3. Application of the KG to drug development, drug repositioning and toxicology research.**
  - a. Showcase the versatility and applicability of the KG in compound centred case studies.

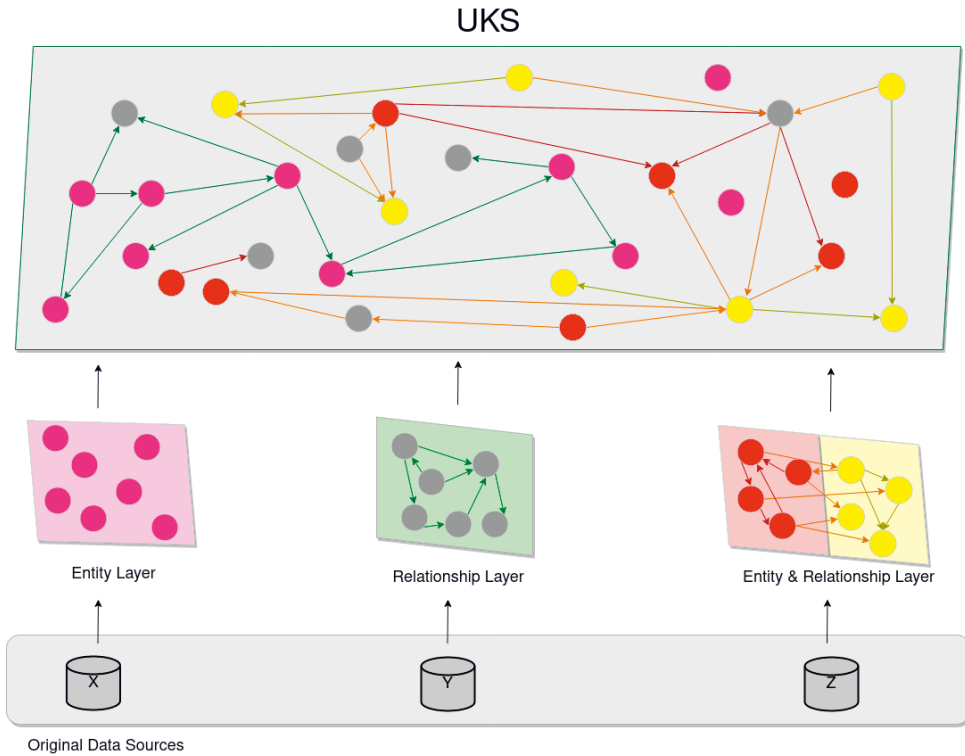


## 7 MATERIALS AND METHODS

### 7.1 The Unified Knowledge Space

The Unified Knowledge Space (UKS) is a bio-medical knowledge graph created to contain a diverse set of publicly available data sets as well as specifically created experimental data, which allows its application in many different areas, such as for drug repositioning, compound discovery, PPI prediction, data retrieval and entity mapping, phenotype grouping, as well as to understand underlying molecular processes in depth. The data types and sources integrated into the UKS are listed in Table A 1.

The different data sources integrated can contribute to the entity layer, indicating that these sources contain information regarding a specific entity only. This data is modelled in the data model as a node and its attributes. The data sources can also contain data with respect to relationships between the same entity type, such as protein protein interactions, which are modelled as relationships in the data model. Lastly data sources can contain information of both layers and between different entity types and/ or different relationship types. Their data is modelled as both nodes and relationships in the later described data model (Figure 10). The different types of information and layers, data sources can contribute to are depicted in Figure 8.



**Figure 8** Describes different types of information and the data layers they can contribute to. Data types only containing entity information (data source X, coloured pink) contain data modelled as nodes and their attributes in the UKS, data sources containing data contributing to the relationship layer (data source Y, coloured green) contain information about the relationships of two entities, which may or may not be of the same type. Lastly data sources can contribute to both layers (data source Z, coloured red and yellow) by containing entity specific information as well as information about the relationships between two entities.

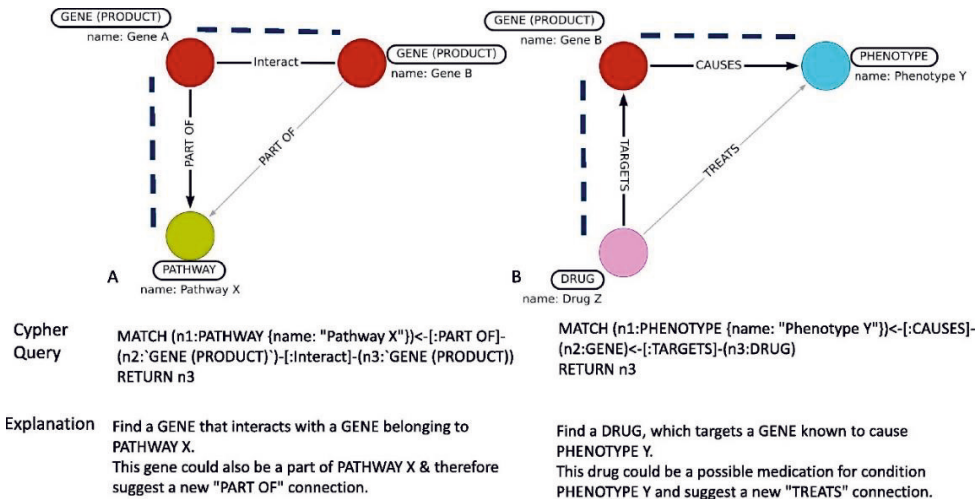
### 7.1.1 Technology

The database management system behind the UKS is Neo4j<sup>4</sup>. Neo4j has been selected over other available database management systems, mainly due to being one of the biggest providers of an open-source graph database system in addition to having a large active community. Additionally, Neo4j provides a “start-up” program, which allows access to the commercial tools and database management system for smaller enterprises and research usage. However, the UKS is of such a scale that the commercial visualization tools are not suited to its size and the database management

<sup>4</sup> neo4j.com

system features coming with the commercial license mostly span enterprise dependent features, such as extensive user and permission management, which are not needed in the way the UKS is deployed for internal research purposes only (Figure A 1). Therefore, the UKS is deployed in the freely available Community edition<sup>5</sup>. The UKS was firstly deployed in Neo4j 3 and upgraded with the development of Neo4j to Neo4j 5 over 4. To start, stop, move and be system independent, the UKS is deployed via Docker Containers (Merkel, 2014), which allows to run the UKS detached from the deployment server, scale hardware usage on demand and move the image between computing infrastructures as needed. The system architecture hosting the UKS is displayed in Figure A 1.

Querying and adding data are performed via Python's py2neo package<sup>6</sup> by sending the Cypher commands directly to the connected Neo4j database. Cypher is Neo4j's query language, which syntax is based on SQL (Structured Query Language) and natively models the graph structure. Examples of Cypher queries are displayed in Figure 9.



**Figure 9** Cypher queries with their graph and textual representation. Figure taken from (Pavel, Saarimäki, et al., 2022).

<sup>5</sup> [neo4j.com/licensing/](https://neo4j.com/licensing/)

<sup>6</sup> <https://py2neo.org/v4/index.html>

## 7.1.2 Data Model

A simplified data model of the UKS is displayed in Figure 10. The data model is created to be flexible, highly adjustable as well as performance orientated for the most common use cases. Therefore, any parameters that are common between multiple entities and are non-numeric are modelled as nodes and the same applies to entities. Each node is labelled with at least one label and sub-labels are attached if possible. These labels can increase node search performance, human readability of the graph and encode information about sub-types or the hierarchical relationship between entities.

Whenever possible numeric data is transformed into categorical data to allow the integration into the graph data model, which allows the matching and re-use of these data-points. For example, patient age is mapped to age groups and expression data is classified for each dataset into expression categories of LOW, MIDDLE, HIGH and NOT EXPRESSED based on the DFP (Discriminant Fuzzy Pattern) method (Glez-Peña et al., 2009). Especially for expression data this allows the re-use and comparability between datasets, since the count values on their own are highly affected by technology and environmental factors (Federico et al., 2020).

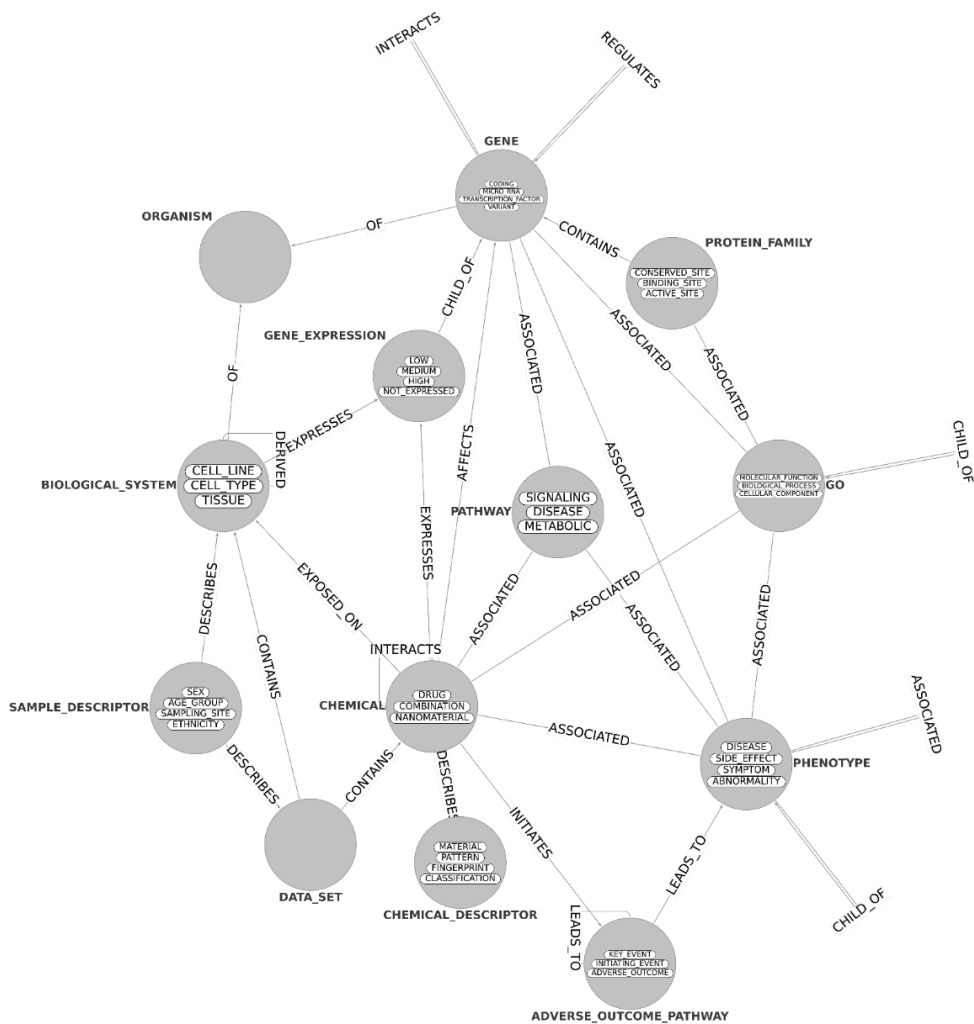
Relationships between entities are modelled as directed edges. Due to technical restrictions of Neo4j each edge must be direct. To avoid adding duplicate edges for bi-directional relationships, each edge is fitted with a Boolean parameter indicating if the edge needs to be interpreted directional or bi-directional. Each edge is of a specific type and no sub-types can be added due to restrictions of the database management system.

Every node and edge are fitted with parameters, which provide at least identifiers of the entity and for relationships the *source* of the relationship. Any further unique edge or node data are added as attributes. When multiple commonly used identifiers exist for an entity, such as gene Ensembl ID (Cunningham et al., 2022), gene Entrez ID (Maglott et al., 2011) and gene symbol, all identifiers are added, when possible, to an entity, while one of the identifiers is selected as the main identifier used in the UKS and the others are mapped towards it. These node identifiers can be used to map between different terms for the same entity when the UKS is used as a knowledge base, as well as simplify the integration of data sources into the UKS by allowing the use of multiple identification systems (when applicable). The edge *source* attribute can be used to signify the strength of the relationship or filter for relevant source data points as well as provides transparency into the origin of the data point.

In addition, the UKS is used to manage experimental meta-data for experimental datasets. Such data is for example gene expression counts (Federico et al., 2020; Kinaret, Serra et al., 2020) or fold changes of an exposure (Federico et al., 2020; Ritchie et al., 2015). This data is stored on a file storage, since these are data points that need to be processed differently depending on the use-case and therefore not suitable to be stored in a graph data model. However, the UKS manages the metadata and points to the relevant file locations, so that search queries for experimental data of specific conditions can be performed within the UKS. The complete data and system infrastructure of the UKS is displayed in Figure A 1.

Node labels and edge types are created custom as needed and if possible existing labels and types are used. Each edge is allowed to be of one type only, while nodes can have multiple labels, as restricted by the database management system. This functionality is used to add additional information or hierarchical order into the nodes, as well as to increase query performance. For example, all genes or gene products are of type GENE, while additional information on the type of gene product is added with additional labels, such as PROTEIN CODEING, PSEUDOGENE or miRNA. Table A 4 and Table A 5 list all node and edge types currently available in the UKS, while Figure 10 showcases different node labels and their sub-labels in a simplified data model.

While the data modelled in the UKS is human centred, it contains information of multiple different organisms, such as common modelling organisms like mouse and rat. Organism specific nodes, such as genes or organism specific pathways are linked to their corresponding organism node, as displayed in Figure 10. For genes, information about the relationships between human genes and a model organisms' genes, such as homologous information are added, which allows to infer knowledge across species.



**Figure 10** Simplified data model of the UKS. In-node labels denote different sub-labels of the main node types, which are displayed in bold outside of the node.

## Technical Attributes

Most node and edge attributes in the UKS contain node and edge specific information extracted from their corresponding data source. However, there are a few identifiers that are used for data transparency, performance, and purely technical reasons. Each node and edge have a unique identifier (*ID*), automatically created by Neo4j, which is used by Neo4j to manage the data, but can be used to identify the same node and edge even if their name has been updated or to improve querying performance in complex queries, due to the natively implemented fast lookup

structure of the *ID* attribute. On creation of each node and edge, a *created* attribute is set on them, which contains the timestamp of its first creation. This, in contrast to the *ID* attribute, is not a default attribute, but instead has been defined in the data model to keep track of creation dates. The already described Boolean *directed* edge property is an attribute defined in the data model, used to describe how the directionality of an edge needs to be interpreted by the user. Due to the technical restrictions of only allowing directed edges in Neo4j, this attribute can be used instead of adding a bi-directional edge twice. Another custom defined attribute used for data transparency is the *source* attribute contained on each edge. This edge attribute is a list and every time an additional source, supporting the same edge is added, its origin source is appended to the *source* list attribute. This attribute preserves transparency of each relationship data point by allowing to trace it back to its original data source(s). In addition, this allows to evaluate an edge based on its source support, which can be an indication of its quality or research interest.

## Indexing to Improve Query Performance

Database indices are data structures stored and maintained by the database management system to improve querying speed by improving lookup. Neo4j 5 introduced different indexing types, from the previously binary search trees<sup>7</sup>. LOOKUP indices on node labels and edge types are natively maintained, while indices for node/edge attributes need to be defined in the data model. Maintaining multiple, especially text-based indices, can increase the database size significantly, since a separate search structure needs to be maintained. Decisions on if an index and what type of index should be created for an attribute in the UKS have been made based on the expectancy of queries on a specific attribute. The offered indices in Neo4j 5 are RANGE index (operate for existence search, equality search, range search, prefix search and list search), LOOKUP index (solve node label and edge type, present by default), (full) TEXT index (operate on strings for contains and suffix searches) and POINT index (only operate on points, not applicable to the data in the UKS)<sup>8</sup>. If multiple indices are available the system will automatically decide which index to use, based on the statement it must solve. TEXT indices are preferred for *CONTAINS* and *ENDS WITH* Cypher statements, POINT indices for distances and in all other cases RANGE indices are selected when available. The default

---

<sup>7</sup> <https://neo4j.com/docs/operations-manual/current/performance/index-configuration/>

<sup>8</sup> <https://neo4j.com/docs/cypher-manual/current/indexes-for-search-performance/>

LOOKUP indices are applied for node label and edge type searches <sup>9</sup>. The created indices, together with their expected search queries are listed in Table A 2.

### 7.1.3 Data Integration

The goal of the data integration process is to link data from different sources, be able to uniquely identify entities as well as to avoid duplicate entries. In the UKS the data integration process focuses on entity recognition, while edges of the same type between the same nodes are updated with the additional supporting sources and edges of different types are added as a new relationship.

Whenever possible, established identification systems are used, however for many entities no widespread identification system exists. In these cases, custom identifiers have been created. For entities where multiple established identification systems exist, one is selected as the main identifier and other identifiers are added as far as possible. The main system is selected based on if there are automatic mapping and search tools or APIs (application programming interface) available as well as based on if an identifier is more commonly used in the local research community. The entity recognition challenge is further complicated by the fact that for many entities there is no complete 1 – 1 mapping between identifiers, such as for genes or phenotypes (chapter 3). In Table A 3 the different entities and the selected identification systems are listed for the main node types in the UKS.

### Entity Mapping

Entity mapping is performed based on a mix of manual identification, external engines and custom developed pipelines. Gene products are mapped to Ensembl Gene IDs via the mygene API (C. Wu et al., 2013). Phenotypes are mapped to NCBI Concept IDs via the NCBI MedGen API (Sayers et al., 2022) based on their names and for terms not found, a broader matching via a custom NLP based pipeline, based on Python's NLTK API (Bird et al., 2009) or created software (Di Lieto et al., 2023) is used. Chemicals are mapped to NCBI PubChem CIDs or SIDs via the NCBI PubChem API (Kim et al., 2023) based on their name or SMILES. Unique sets, such as pathways (Jassal et al., 2020; Kanehisa et al., 2017; Martens et al., 2021), Gene

---

<sup>9</sup> <https://neo4j.com/docs/cypher-manual/current/indexes-for-search-performance/>



Ontology terms (Ashburner et al., 2000; The Gene Ontology Consortium, 2021), gene sets (Liberzon et al., 2015) or Adverse Outcome Pathways<sup>10</sup> (Saarimäki, Morikka, et al., 2023) are identified by their origin system identifiers, since they are unique sets created by the source system. Non standardized terms, such as tissues and cell types are identified with custom defined naming.

Due to the complexity and non-standardized way of entity identification in the life sciences, the entity identification process is not 100% and some entities cannot be identified or are mapped to the wrong entity. When identified, errors in the mapping are corrected manually and entities not mappable are discarded or added manually based on their origin and quality. Further users can submit possible detected errors in the UKS, which are manually corrected if confirmed. Possible suspect edges or entities are flagged in the UKS and a description is added, so that other users are informed and when more information is available a decision about the quality can be made. For example a phenotype, identified by an established ID system, but cannot be mapped to a term in NCBI MedGen (Sayers et al., 2022) will be added based on its name and alternative ID, if the alternative ID is an established identification system, such as Human Phenotype Ontology (Köhler et al., 2021) terms or KEGG disease (Kanehisa et al., 2017) terms, since it is assumed that the entity can be found and mapped in the future to other sources. However, a phenotype, identified by a custom name created by the source author will be discarded, since it cannot be linked to other sources or information and therefore on a large scale will not add information to the UKS.

## Version Control

When updated versions of already integrated data sources are added to the UKS, their old data is not removed but rather the new version is added as a “different source”. This means that for example on a KEGG (Kanehisa et al., 2017) update, any relationships added with the current KEGG version are added as any other new data source and the source identifier, added on each relationship, reflects the KEGG version. This allows the user to filter their queries based on which version they are interested in, but also to retrieve older identifiers and associations. For example, GO (The Gene Ontology Consortium, 2021) terms may be removed or updated in a new release, however there are many other data available which use GO identifiers or link

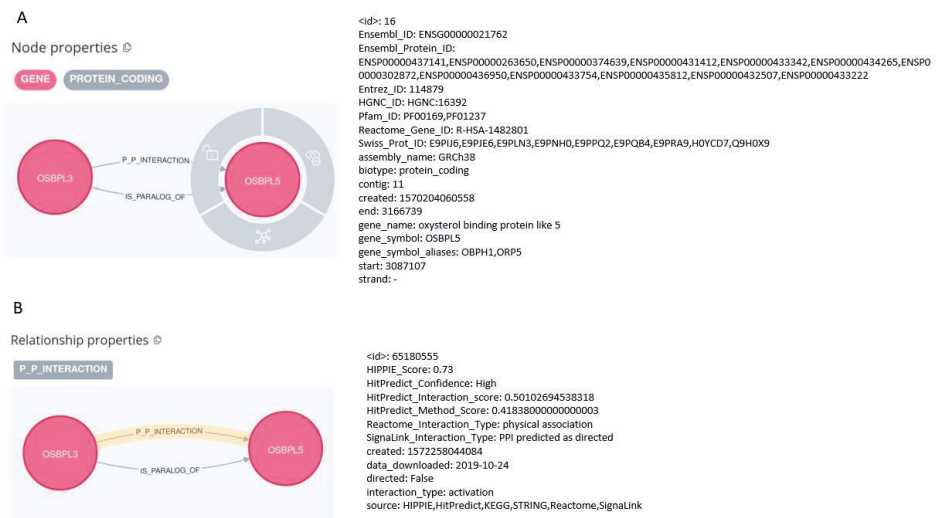
---

<sup>10</sup> aopwiki.org

knowledge to these identifiers. These external sources may not be updated at the same time, if they are. So, by removing non-current identifiers from the UKS, the UKS would lose any other information linked to these identifiers. By adding version information instead, it allows to retrieve old and new information at the same time, while the user can evaluate the version information and make decisions based on available data and their use-case.

## 7.1.4 Data Retrieval

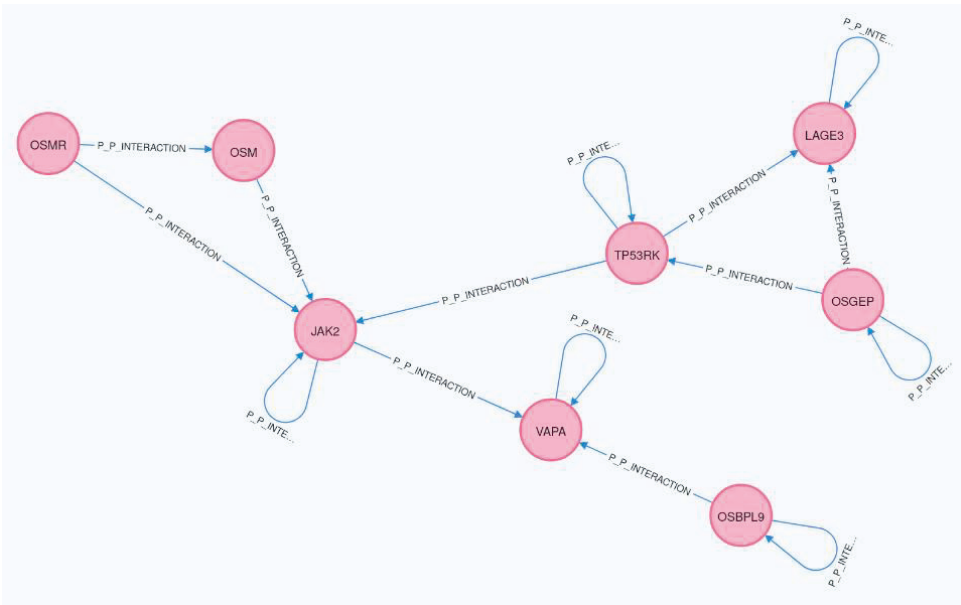
Data points can be retrieved from the UKS as sub-networks of the UKS, which can be networks directly contained in the UKS or inferred from the UKS. Individual data points can be queried based on relationships between entities and in addition entity specific information can be retrieved from the node attributes. These data can be used to describe an entity or relationship further but is not linked to a different entity in the UKS (Figure 11A). Such data points are for example alternative node identifiers or the edge *source* attribute from which the origin of each relationship can be traced (Figure 11B).



**Figure 11** A) Example of a GENE entity node in the UKS and some of its attributes, which provide further information about the node. B) Example of a P\_P\_INTERACTION edge in the UKS and some of its attributes.

## Robust Single Layer Network

A single layer network only contains one type of data between two entities, which may or may not be of the same type. The relationships between the entities can come from one data source or multiple data sources. When multiple data sources are available, this information can be used to extract robust networks. An example of a single layer network would be the protein protein interaction data layer in the UKS (Figure 12).

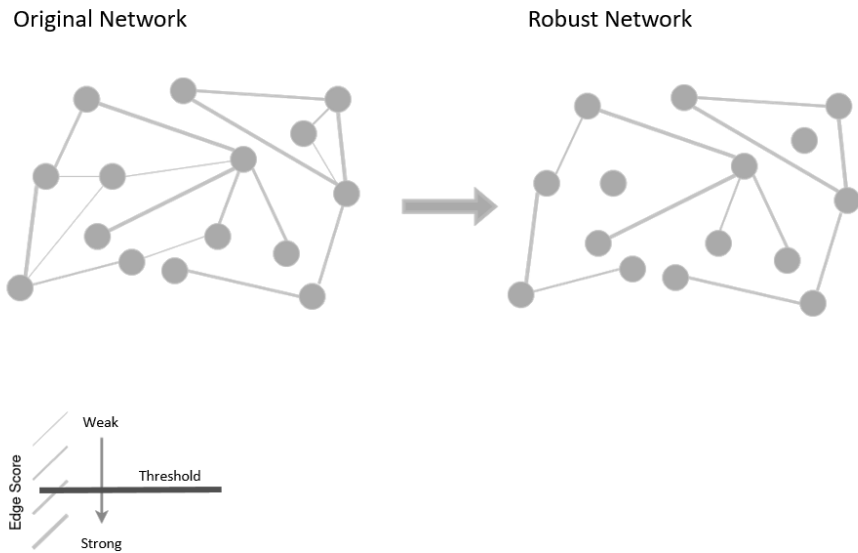


**Figure 12** An example of a homogenous node and edge network, showcasing the protein protein interaction layer in the UKS.

### Source Support Global Threshold

Robust edges can be extracted based on a global threshold. This threshold states how many sources need to support a data point (relationship) for it to be considered “TRUE” and to be kept in the final network (Figure 13).

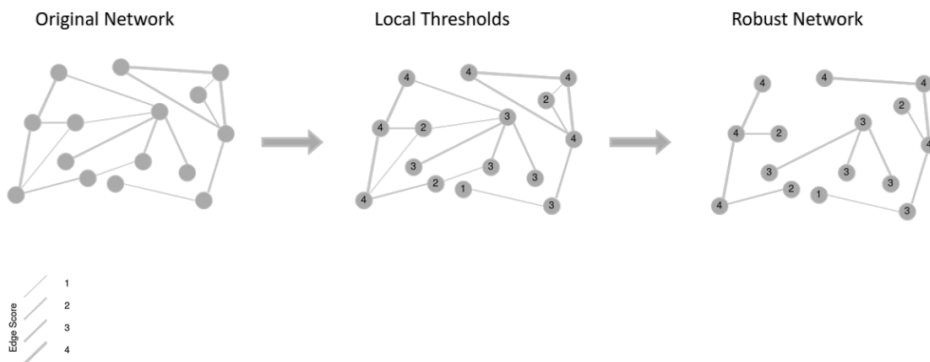
This method works well when data source availability is equally distributed across all entities. In case there is a data availability bias, this method may enforce the bias further by discriminating against entities that have less data available or are less researched in general. In this case a local threshold may be more appropriate.



**Figure 13** Extraction of a robust network from the UKS based on a global threshold. The edge weight indicates its strength of the relationships.

### Source Support Local Threshold

When a local threshold is applied to cut edges, a source support score is calculated for each node independently. This can for example be the mean, median, minimum, or maximum number of sources supporting all edges of a node. This method reduces data availability/ research bias, since for each entity the score (number of sources to be considered a “TRUE” edge) is determined independently (Figure 14).



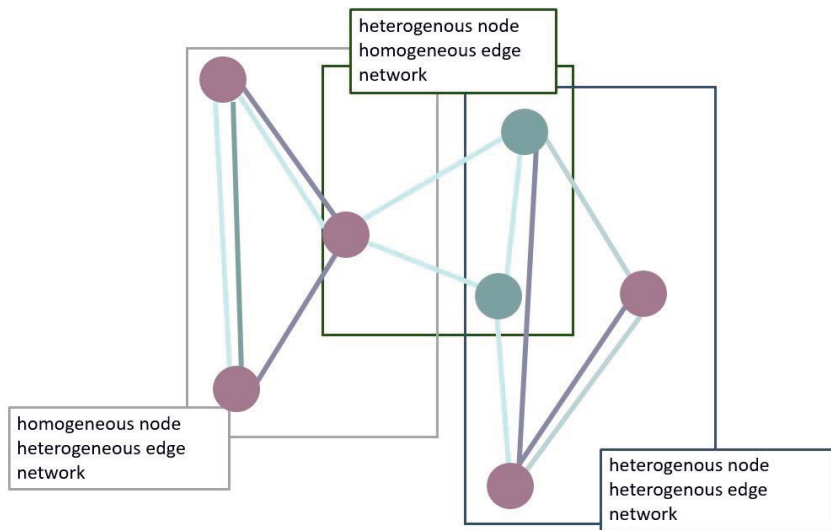
**Figure 14** A robust network is extracted based on local (per node) estimated edge weight thresholds. In the example for each node only the edges with at least the maximum edge weight of all the edges connecting to that node are considered. An edge is kept if this is true for at least one of the nodes connected by the edge. The values in the nodes indicate the computed local edge weight threshold for that node.

## Multi-dimensional Network

A multidimensional network combines information of different data types, which can be in a homogeneous or heterogeneous node network. The extracted multilayer network can also be transformed from a heterogeneous to a homogeneous network.

## Homogenous Network

A homogeneous multilayer network contains different relationships between one type of entity (Figure 15). Many network analysis algorithms expect only one type of edge between entities, so it may be advisable to merge the different layers into a single layer. This can be done by creating for each data layer a weighted network, where the score represents either a defined edge weight or for example the source support score of the edge. The single layer networks can either be directly cut as described previously or cut after the individual layers have been merged. Cutting after merging allows to keep edges that may not have made the cut in the individual networks but when combined are scoring high enough to be considered. Which strategy to employ depends on the analysis to be performed. The individual layers can be merged by either summing, averaging (or any other metric) the single layer scores or using the number of layers present as total score. The resulting network can be cut globally or locally (section 7.1.4), depending on the data and planned analysis.



**Figure 15** Example of a multilayer network in the UKS, combining multiple edge types and node types. Grey) a network of the same node type connected by different edge types. Green) a network connecting different node types by the same edge type. Blue) a network combining different node types connected by different edge types. Entity and relationship types are indicated by their node and edge colour respectively.

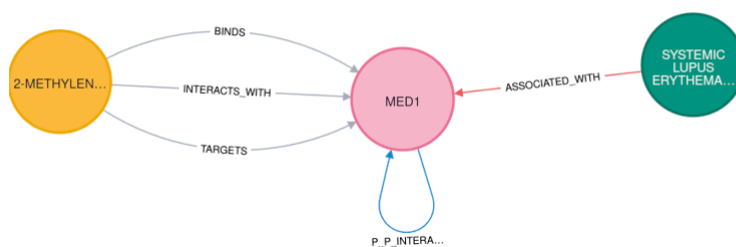
### Heterogenous Network

A heterogenous multilayer network contains (different) relationships between different entity types (Figure 15). If from a heterogeneous multilayer network, a robust heterogenous multilayer network needs to be extracted, the same merging and cutting methods as described earlier can be deployed. In case the heterogenous network information needs to be transformed into a homogenous network, the to be extracted edges are not directly contained in the data but are calculated based on for example paths in the graph. So can for example edges be created between entities of the same type based on how many common neighbours of a different type they have, or what the shortest path distance between them is in the graph or a specific data layer. Once the heterogenous network has been transformed into a homogenous network the same merging and cutting principles as described earlier can be applied. The same can be applied when multiple heterogenous networks are transformed into a homogenous network in order to be merged.

## Path Based

Relationships can be directly retrieved from the UKS to provide information to which other entities a specific entity is linked. Data points can be retrieved from the UKS, that are not directly the result of individual data sources but only visible through their representation as a graph. These data points can be retrieved based on meta-paths, which are paths in the network where the to be visited edge/ node (types) are defined. Figure 16 displays a Cypher query stating the node and edge types to be visited to retrieve a possible drug for the treatment of a disease by specifying that the drug target needs to be associated to the disease.

```
MATCH p=(DRUG)-[:TARGETS]->:(GENE)-[:ASSOCIATED_WITH]-(:DISEASE) RETURN p LIMIT 1
```



**Figure 16** Path based data extraction in the UKS. The example showcases a Cypher query stating the node and edge types to be visited to retrieve a possible drug for the treatment of a disease by specifying that the drug target needs to be associated to the disease.

## Entity Based

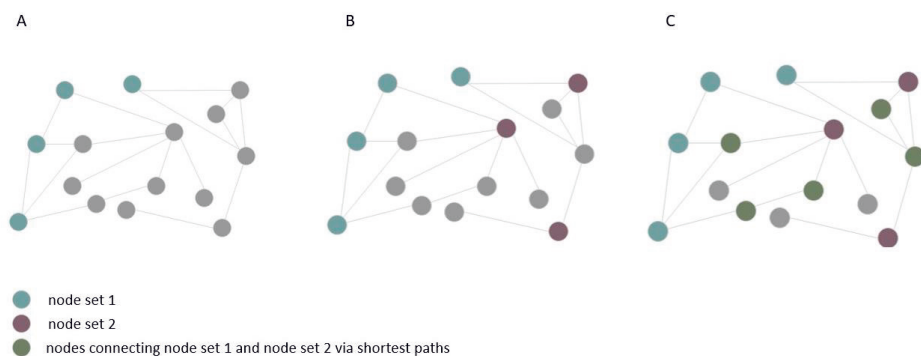
Entity based data extraction, focuses on the data stored in an entity's node properties. These are datapoints only applicable to the individual entity and not in relationship with other entities. Such data can for example be alternative names or identifiers. These queries are especially helpful when a mapping engine between different identification systems is needed, such as from Ensembl Gene IDs to gene symbols or additional entity information is requested, such as SMILES for a chemical compound (Figure 11A).

## 7.2 Intermediate Genes – Identifying Relevant Non-Measured Genes

Transcriptomics or differential gene expression analysis only allows to identify relevant gene products at a specific point in time. Therefore, not all gene (products) taking part in a molecular process can be identified. PPI networks, gene regulation networks or similar gene gene networks contain possible cascades of involved gene (products) in a molecular process. With network analytical methods these hidden intermediate genes can be identified.

In publication I all shortest paths between two sets of genes are computed with the Python NetworkX API (Hagberg et al., 2008). Paths of size 1, i.e., where there exists a direct link between two genes of the different sets are excluded, so that only paths with at least one connecting intermediate gene are considered. To determine statistical significance of the newly identified intermediate genes a hypergeometric test is performed, comparing the frequency of the gene being an intermediate gene between all shortest paths in the gene network against the frequency of being on the shortest paths between all gene pairs of the two compared sets. The hypergeometric test is performed with Python's SciPy API (Virtanen et al., 2020) and the p-values are corrected for multiple testing with the Benjamini Hochberg method (Benjamini & Hochberg, 1995) from the statsmodels Python API (Seabold & Perktold, 2010). Figure 17 showcases the shortest path based intermediate gene identification method. This metric can be performed on any gene network of relevance to the research question and can further be applied to other homogeneous or heterogeneous networks.





**Figure 17** Shows how intermediate nodes (genes) can be identified on a network via shortest paths. A) Node set 1, can for example be genes known to directly interact with a compound. B) Node set 2, can for example be genes measured as differential expressed in an exposure study. C) Nodes on the shortest paths between node set 1 and node set 2, this can for example be genes propagating the signal from the directly targeted genes to the measured genes. These genes can provide further insight into the molecular processes taking place.

### 7.3 VOLTA – Network Analytics & Multi Network Clustering

Publication II VOLTA is a Python package, providing necessary functions for the analysis and comparison of biological networks, with a focus on gene co-expression networks (Marwah et al., 2018; Pavel, Serra, et al., 2022). The aim is to provide a highly flexible and modifiable package, which in contrast to comparable resources is not limited by a graphical interface, which makes it easily useable for other types of analysis or networks, such as can be extracted from the UKS. For each possible analytical step, VOLTA tries to provide different available algorithms and consensus strategies (where applicable) to allow users to fully customize their analysis. By building on basic Python objects and established Python APIs (such as NetworkX (Hagberg et al., 2008)) for function input and output, the user can further add their own algorithms or analysis steps in-between, without needing to perform complex transformations of the input and output object types.

The aim of VOLTA is not to re-implement algorithms for which already established Python APIs exist, but rather to make use of their open-source and community availability and integrate their API calls into VOLTA. Algorithms without a suitable implementation are implemented based on their corresponding publication. In

addition, VOLTA also provides ready-made analysis pipelines, which can easily be modified or used as is, to allow the usage of VOLTA through inexperienced users.

### 7.3.1 Comparison of Drug MOA via Network Analysis

Gene expression data is retrieved from the Lincs 1000 Gene Expression Omnibus (GEO: GSE70138). The gene co-expression networks are computed with INFORM (Marwah et al., 2018), with the selected inference algorithms of CLR (Faith et al., 2007), ARACNE (Margolin et al., 2006) and MRNET (Meyer et al., 2007, 2008), based on all available correlation and mutual information metrics as implemented in the R minet package (Meyer et al., 2008). For the case study the A549 cell line treated with dasatinib and mitoxantrone at 10  $\mu$ m and 24 hours are selected.

#### Differential Centrality Analysis

Different node centrality metrics are calculated for each network, the nodes are ranked for each individual centrality measure (degree, closeness and betweenness centrality) and a median rank across all centrality metrics for each node is estimated with `volta.distances.node_edge_similarities.sort_node_list()`. The rank difference for the same node (gene) between two networks is estimated and the nodes are ranked by their difference in ranks. This allows to identify genes, which network position and therefore their importance in the gene network, are strongly altered by the condition(s) under investigation.

#### Comparing a Nodes Neighbourhood

Random walks are a method often employed to characterize local network structures. In densely connected structures, random walks will rarely leave the neighbourhood while in looser structures the walks are more likely to explore further distant areas. Characterizing a nodes neighbourhood allows to identify genes likely connected to the same molecular function. When comparing the change in neighbourhood between two networks/ conditions, this can help in identifying if the same molecular processes or different ones take place.

For each node in the network  $10 * \text{node degree}$  random walks of length five are computed with `volta.example_pipeline_wrappers.get_walk_distances.helper_walks()`. The large number of short walks increases the accuracy of characterizing the explored

nodes neighbourhood. For each starting node the visited nodes are counted and ranked based on their occurrence with *volta.example\_pipeline\_wrappers.get\_walk\_distances.helper\_get\_counts()*. To compare the two networks a Kendall rank correlation for each node between its ranked visited nodes across the networks is estimated.

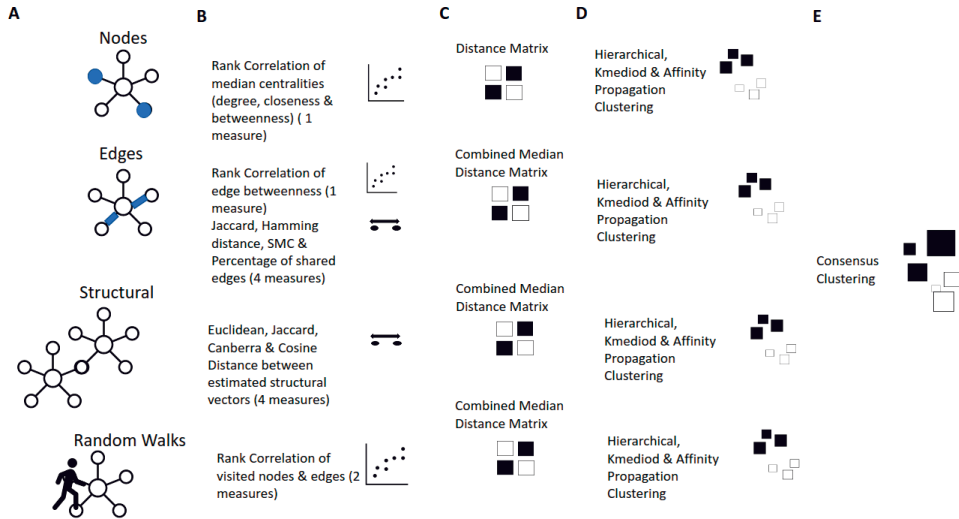
## Comparing Networks via Their Communities

Different algorithms perform differently, extract communities based on different parameters and in result the community structures for the same network can vary strongly. While there are metrics to estimate how “good” an identified community is, there is no established algorithm which is considered best on biological networks in general. Therefore, VOLTA provides a consensus strategy, which takes as input different community partitionings and computes a consensus. The consensus partitioning is computed with *volta.communities.fast\_consensus()* and the individual communities are functionally enriched via the Panther Enrichment API (Mi et al., 2021) for Reactome pathways (Jassal et al., 2020).

### 7.3.2 Grouping Networks Based on Network Metrics

The networks are created the same way as described in the previous section. For this case study 20 cell lines, treated with 10  $\mu\text{m}$  dasatinib for 24 hours are selected. The available 20 cell lines are listed in Figure 26.

The networks are clustered based on their node similarities, edge similarities, topological similarities and subgraph similarities. The whole pipeline is displayed in Figure 18.



**Figure 18** Clustering pipeline available in VOLTA. Image taken from publication II.

## Node Similarities

For each network their node centrality scores are estimated and the Kendall Rank correlation between the ranked nodes for each network pair is computed with `volta.example_pipeline_wrappers.get_node_similarity.estimate_similarities_nodes()`. The so resulting correlation matrix is transformed into a distance matrix with distance  $d = (1 - correlation) / 2$ . Since the investigated networks are all made of the same nodes, distance metrics that are based on the number of shared nodes are not computed.

## Edge Similarities

Next the similarity for each network pair is computed for their edges, based on edge betweenness scores and their overall overlap of edges. The investigated networks are binary, so all edges are of equal weight. With `volta.example_pipeline_wrappers.get_edge_similarity.estimate_similarities_edges()` the Jaccard distance (Jaccard, 1908), Hamming distance (Hamming, 1980), simple matching coefficient distance as well as the fraction of edge overlap between each network pair is computed. In addition, the Kendall Rank Correlation based on the edge betweenness scores for the top 100 edges between each network pair is calculated. Non distance matrices are again transformed into a distance matrix and if necessary,

scaled to be in [0,1]. The individual distance matrices are combined into a consensus matrix with `volta.clustering.create_median_matrix()`.

## Topological Similarities

For each network a feature vector based on different topological measures is computed with `volta.example_pipeline_wrappers.get_network_structural_vector.estimate_vector()`, which computes graph radius, diameter, number of nodes and edges, density, average clustering, fraction of existing and non-existing edges, number of cycles, cycle size distribution, shortest path distribution, clustering coefficient, degree distribution, closeness centrality distribution and betweenness centrality distribution. For each network pair the Euclidean, Canberra, correlation, cosine and Jaccard distance are estimated with `volta.example_pipeline_wrappers.get_network_structural_vector.matrix_from_vector()` and combined into a single distance matrix with `volta.clustering.create_median_matrix()`.

## Substructure Similarities

For each node random walks are performed, and their neighbourhoods compared as described in section 7.3.1 (Comparing a Nodes Neighbourhood) with walk length five and  $3 \times \text{node degree}$  walks per node. For each network pair the visited nodes and edges for the same starting nodes are ranked by their occurrence and the Kendall Rank correlation is estimated. The resulting matrices are transformed into a distance matrix and combined into a consensus matrix.

## Clustering

On each of the four resulting distance matrices K-mediod (`volta.clustering.kmedoids_clustering()`), Affinity propagation (`volta.clustering.affinityPropagation_clustering()`) and hierarchical clustering (`volta.clustering.hierarchical_clustering()`) are performed. Where applicable parameter selection is performed with `volta.clustering.multiobjective()`, which is set to prioritize within cluster similarity, maximize between cluster dissimilarity and favour an even cluster size distribution. For the selected parameters, algorithms relying on randomness are run 10 times, non-randomness-based algorithms are considered 10 times to avoid biasing the final consensus clustering towards one algorithm. The

consensus clustering is computed with *volta.clustering.consensus\_clustering()*, which builds an agreement graph between the individual cluster groupings, removes weak edges and performs community detection  $x$  ( $x=10$ ) times. This process is repeated until it converges, or the maximum number of iterations is reached. Weak edges can either be identified based on a set threshold or automatically be determined based on permutations of the adjacency matrix.

## 7.4 Grouping of MOA Across Systems and Data Sets

Publication III KNeMAP is a method, available as a Python module, which is aimed at comparing and clustering gene expression data. Due to its dependence on a multi-dimensional and expression data independent prior network, it is more robust to noise in comparison to alternative methods and can be used across data sets and technologies.

### 7.4.1 Extraction of a Homogenous Multilayer Network

The aim of the prior network is to be an independent robust data source, describing how similar genes are across different data types. For this individual data layers between gene entities in the UKS are extracted, in addition to multidimensional heterogenous networks, which are transformed into homogenous networks, as described in section 7.1.4 (Multi-dimensional Network). Each individual network is a weighted network. The networks resulting from a single layer in the UKS are weighted based on their source support score (section 7.1.4 (Robust Single Layer Network)), the generated homogenous networks from the heterogenous networks are weighted based on the number of common neighbours two entities share. To avoid biasing the final prior network by the availability of similar individual networks, the individual networks are first evaluated based on their similarity of existing binary edges. The distance matrix between the individual data layers is computed with VOLTA (publication II) based on a Jaccard distance, Simple Matching Coefficient and the fraction of shared edges. The individual distance matrices are combined into a single distance matrix as described in section 7.3.2. Hierarchical clustering is performed on the final distance matrix with SciPy's python API (Virtanen et al., 2020). The individual layers in the so detected clusters are merged first. Each network in a cluster is scaled to all have the same median edge weight and summed up. The same process is performed between the so resulting cluster networks, which results in the final multilayer prior network.

## 7.4.2 Detection of Similar Genes

The prior network is partitioned with the publication II VOLTA API (*volta.communities.agglomerative*). The agglomerative network partitioning algorithms has been selected due to its ability to yield multiple small communities, which in this case are preferred to keep the specificity of the grouped genes.

## 7.4.3 KNeMAP Vector

Publication III KNeMAP maps the  $x$  most deregulated genes of an exposure against the prior network communities. For the case study  $x=200$  is selected, which keeps the specificity of an exposure but allows to identify similarities between them. For each community, the fraction of deregulated genes falling into it, is computed and the scores for all communities are combined into a feature vector describing the exposure.

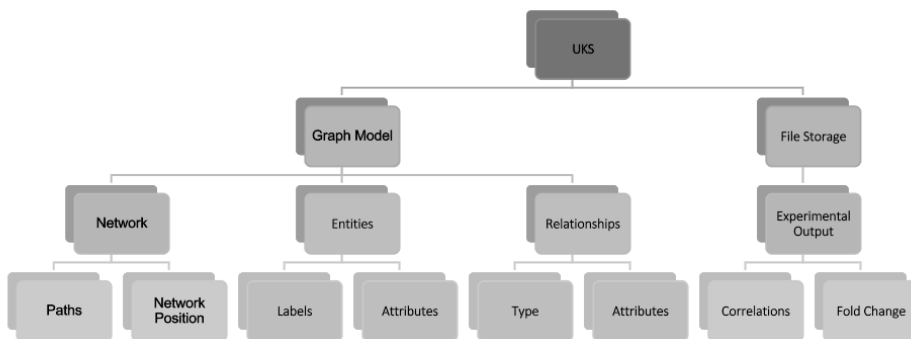
## Noise Analysis

To showcase the noise robustness of the KNeMAP vector against other metrics, different levels of noise (0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1) are added to the raw gene expression values. In a second experiment the selected  $x$  most, deregulated genes are perturbed with the same noise levels. For each noise level the difference between the newly created vector and the baseline vector (noise = 0) is measured across all available exposures. The median distance is computed, and the difference measured via the area under the curve (AUC).

## 8 RESULTS

### 8.1 The UKS as a Robust Multidimensional Data Source for Data Retrieval, Knowledge Linkage and to Support Transcriptomic Analysis

The UKS is an integrated data source, comprising to date 83 independent data sources, 13.3 million nodes and 1.02 billion edges. Each node and edge contain additional data points unique to the entity or relationship. This makes the UKS comprise approximate 3.3 billion data points in total, which can directly be accessed via nodes and edges (Table A 4 and Table A 5). In addition, approximately 600 million data points (Table A 6) are stored in the file storage (Figure A 1) managed by the UKS. The data point estimate, however, does not include data points stored indirectly in the graph via paths or hidden links, hierarchical node relationships contained on the node labels, as well as the data that can be extracted by viewing an individual entity with respect to the whole system or the surrounding entities, such as node centrality measures or network communities (chapter 4). These dimensions add another multiple billion indirectly modelled data points to the total number of data points in the UKS. The dimensionality of data points stored in the UKS is showcased in Figure 19.



**Figure 19** Dimensionality of data points stored in the UKS.

Data integration is performed successfully for the 83 data sources. The data integration tasks consist mainly out of two main tasks, entity identification and



mapping as well as relationship classification. Entity identification and mapping is performed mainly via external APIs, as listed in Table A 3. Gene (product) entities are mapped between identifiers, such as gene symbols, Entrez Gene ID (Maglott et al., 2011) or Ensembl Gene ID (Cunningham et al., 2022) via the mygene API (C. Wu et al., 2013). The usage of Ensembl Gene ID allows to add the exact loci of each gene as additional attributes to the entity in the UKS, making the entity, while identified by the assigned ID, independent from its name but based on its physical location. This contrasts with for example gene symbols. In addition, the Ensembl platform provides itself an easy to access API, which provides additional information about genes, such as location or alternative identifiers, which are all data points added to the entity nodes in the UKS. In the UKS all gene (products) are modelled as gene nodes, identified by their Ensembl Gene ID, meaning that for example proteins are mapped to their corresponding gene. This decision has been made to simplify the data integration problem but implies that for example protein protein interaction data is mapped to the corresponding genes and some information is lost during the integration stage. However, the data model would allow to add protein nodes as separate node types if this is required in the future. The usage of established ID systems for gene (product) information is wide-spread and therefore most entities can be integrated into the UKS. Further there is not always a 1-1 mapping between gene identifiers available, in such cases multiple data points are added to the UKS / multiple data points are merged to one based on the corresponding Ensembl Gene ID, which has been selected as the UKS gene (product) identification system. For example, if protein protein interaction data is integrated, which contains interactions between proteins which are all mapped to the same GENE entity in the UKS then a self-edge is added to the UKS but the detailed information interactions between the proteins is simplified in the UKS representation.

Chemical compounds are identified by their PubChem CID or SID (Kim et al., 2023) and are mapped to these via the PubChem API either based on the name provided by the source system or their chemical structure represented as SMILES. PubChem is one of the largest chemical databases available and has been selected as the UKS reference system for chemicals and drugs due to its easy to use and highly accessible API, which allows the automation of the integration task, as well as provides additional information about entities, such as alternative identifiers or PubChem fingerprints. Chemical substructures are identified and mapped based on their SMILES representation. Well defined compounds, such as approved drugs are in most source systems easy to map via the PubChem API, even when represented with a name instead of a standardized identification system or their structural information. As a secondary system for drugs, DrugBank IDs are used and for

compounds which are not in PubChem, such as engineered nanomaterials, custom naming as well as physical properties are used, since for these compounds not only the chemical makeup but also its engineered structure matters for their identification. The usage of PubChem IDs and the definition as drugs as a sub-label of chemicals, indicates that drugs of different manufactures but the same core-chemical composition will be added as the same CHEMICAL entity in the UKS and therefore some manufacturing or country specific information may get lost in the integration task. Most compounds are identifiable based on this method, however for some entities no match can be found, which often is a result of custom compounds used in a source system. Such compounds are often not available or a result of small modifications of existing compounds. If a compound can't be identified it is discarded based on the assumption that (custom) compounds are unlikely to appear in future source systems and therefore are not linked to future information, which does not contribute to the information in the UKS on a global scale.

Phenotypes are identified in the UKS based on their NCBI MedGen Concept ID (Sayers et al., 2022), which are when possible mapped via the NCBI MedGen API. MedGen, similar as PubChem has been selected due to its easy-to-use API, which can translate most similar phenotype terms to an entry in its database. Additionally, it is a large database for which the API also returns additional information about a Phenotype, such as alternative IDs, such as Orphanet IDs or HPO IDs (Köhler et al., 2021) . Phenotypes are, out of the entities mainly identified via external APIs, the most challenging to integrate and the entity type experiencing the most loss. The main reason for this is that Phenotypes are usually reported as text, as well as that the classification granularity and disease naming is highly subjective and language dependent. Therefore, this is the only entity type where if no direct match can be found the phenotypes are mapped to the most likely match(es) via a custom NLP pipeline (section 7.1.3) in order to minimize data loss, which is for the other entity types so minimal that it can be neglected or in the most cases it happens due to the deprecation of the by the source system used identifier. If a phenotype can be mapped to multiple Concept IDs multiple data points are added to the UKS, while if multiple source system phenotype entities are mapped to the same Concept ID the information is merged onto the specific node in the UKS. Further PHENOTYPE nodes can have additional sub-labels such as DISEASE, SYMPTOM or SIDE\_EFFECT, which are again highly subjective to the source-system author and in multiple cases depend on a specific case, so can headache be the side effect of a medication, the symptom of a disease or be considered a disease itself. Therefore, a label does not exclude the addition of other labels and only indicates that this

phenotype may be considered as any of these subcategories under specific circumstances, but it does not necessarily need to apply in all cases.

Organism nodes are identified by their NCBI Taxonomy ID and due to the limited species considered in the UKS these mappings can be done manually. Pathways, AOP, GO (The Gene Ontology Consortium, 2021) and similar entities are identified by their source system identifiers in the UKS together with their names. These entities are commonly reported in third party source systems by their official IDs and therefore integration into the UKS can be done easily and entity loss is very rarely observed. The few cases an ID cannot be identified in the UKS as well as when queried on the source reference system, mostly appear, when a deprecated identifier is used in the source system which has been deprecated before the first version of the entities have been integrated into the UKS. In these instances, it has been decided to not include the datapoint due to its deprecation and the assumption that it likely will not appear in other future datasets to be integrated. Tissues, cell types and cell lines are identified by manual created naming, since a widespread identification system is not available. Further there is no clear definition when an entity is a tissue, organ, system or cell type and the definition is subjective. Therefore, a manual standard for the UKS needs to be created. Additionally, many experimental datasets will report tissues only but then maybe contains a few entities that are under a different definition not considered a tissue (e.g. the GTEx dataset (GTEx Consortium, 2013), which leads to mismatches between source systems and entities in the UKS. Since TISSUE and CELL\_TYPE nodes are kept separate in the UKS this can also lead to duplicate entries due to different entity type definitions of the source systems. Therefore, manual curation and entity definition is needed for these entity types, and the applied semi-automatic entity recognition and mapping is performed via custom software (Di Lieto et al., 2023) , due to the lack of suitable APIs and entity definition.

The data engineering tasks, however, does not always allow for a 1-1 translation of source system to UKS data model. In case multiple UKS entities can be mapped to a source system entity, the source system information is added to all matched UKS entities. On the other hand, if multiple source system entities can be matched onto a single UKS entity the source system information is merged onto the individual UKS entity. Entities may also be discarded if it is not possible to retrieve a possible match against the UKS entity system. This can happen if a source system uses deprecated identifiers that may have been removed from current (used) versions of the identifier source system. For example, if a data source report GO identifiers that have been deprecated before any version integrated into the UKS they cannot be

matched and therefore the data point will be discarded. Section 7.1.3 discusses how identifiers are handled that have been deprecated after their integration into the UKS. In summary the entity recognition and integration task can be mostly automated by relying on external APIs but require the selection of a reference system and consistency in its application. Further it needs to be acknowledged that 100% data translation is not possible, and, in some cases, granularity will be added in the UKS representation, while in others it is simplified. However, the goal of the Big Data integration is, that large scale data is available in a unified format, and the drawback of mismatched data points or missing data points can be neglected in comparison to the available data. Without the simplification of the problem, relying on external source system identifiers and mapping technologies the large-scale data availability of the UKS would not be possible in the limited amount of time and its maintainability as well as expandability could only be achieved with enormous manual effort. The selected approach for entity mapping and identification has in conclusion worked well when weighted between its resource intensity (time, manual effort) and the observed data loss/ data simplification.

Relationship classification is done manually by understanding the data provided in the individual data sources and their information type being mapped to existing relationship types if possible. If a new type of relationship is introduced a new edge type is defined in the data model. Table A 5 lists the currently defined edge types in the UKS. Also, here a decision between simplification, i.e., using existing relationship types and granularity needs to be made. New relationship types are only introduced if no existing type can meet its needs and if the new type would provide significant additional information or can be expected in the future from different source systems. If this is not the case the general ASSOCIATED\_WITH relationship type is used.

The to date existing node types, their sub-labels and most common attributes are listed in Table A 4. Table A 5 showcases the same for the to date existing edge types. This data is used to estimate the currently available data points in the UKS. Due to the graph data model, data points are distributed across multiple dimensions (Figure 19 and Figure 8). Node and edge labels add information about the entities and relationship types. Especially for nodes, the sub-labels contain information about the sub-types of entities, such as if a GENE is PROTEIN\_CODEING or if a PHENOTYPE is considered a DISEASE (Figure 10). Node and edge attributes are data points of a specific entity or relationship, however not all nodes and edges of the same type have the same attributes. The nodes and edges itself comprise data points describing the relationships between entities. Further not only directly existing

edges but also paths and hidden links in the network model can be seen as data points. Due to the complexity, the available data points cannot be exactly calculated. In the estimate the number of entity data points is estimated for each main node label \* the number of its most common attributes. For relationships a similar approach is taken by computing the number of edges of a specific type \* the number of its most common attributes. The data point estimates are listed in Table A 4 and Table A 5.

### 8.1.1 Robust, Multi-Source Support Data

The UKS can be used as an integrated data source, where multiple sources for the same data points are combined to infer robust data points as well as to minimize biases resulting from considering single data sources only (section 7.1.4). In publication I the UKS has been used to create a robust PPI network across four independent data sources, by only keeping edges that are supported by 75% of the available data sources (at point of publication) (HIPPIE (Alanis-Lobato et al., 2017), KEGG (Kanehisa et al., 2017), HitPredict (López et al., 2015) and STRING (Szkłarczyk et al., 2019)). The final network contained 20,793 nodes, 132,244 edges with a network density of 0.0006.

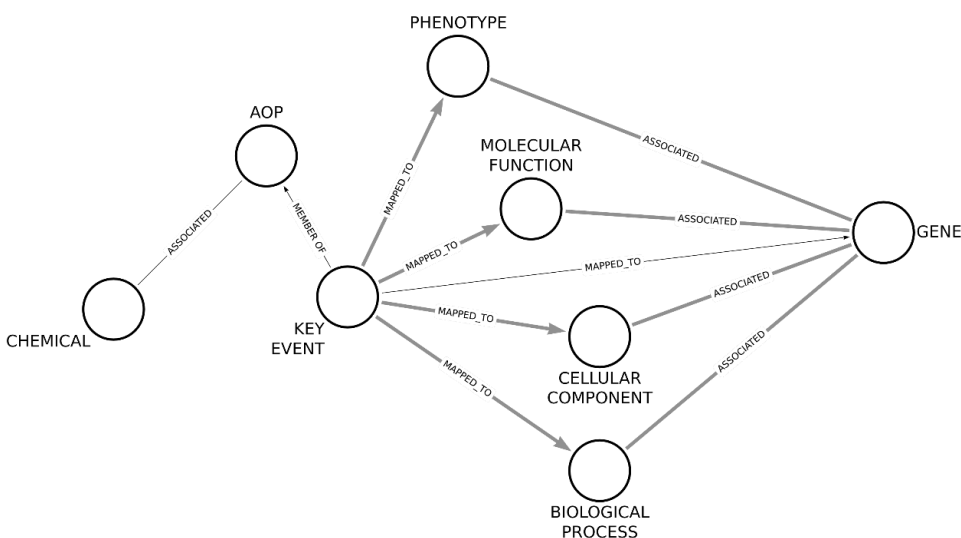
### 8.1.2 Multidimensional Data

In publication III the UKS is used to construct a multidimensional gene gene network, which is comprised out of 11 data types and 12 data sources, as described in section 7.1.4 (Multi-dimensional Network). The final prior network used in publication III contains 22,316 gene (product) nodes, connected by 213,784,257 edges. This network is used as an independent prior to reduce noise in expression data sets, minimize differences between biological system exposures as well as to allow the direct comparison between different expression data sets.

In Federico et al. (Federico, Fratello, et al., 2022) a similar multidimensional gene gene network is constructed from the UKS, comprised out of 6 different data types and 12 data sources, used to evaluate the coverage of different drug combinations for the investigated phenotype.

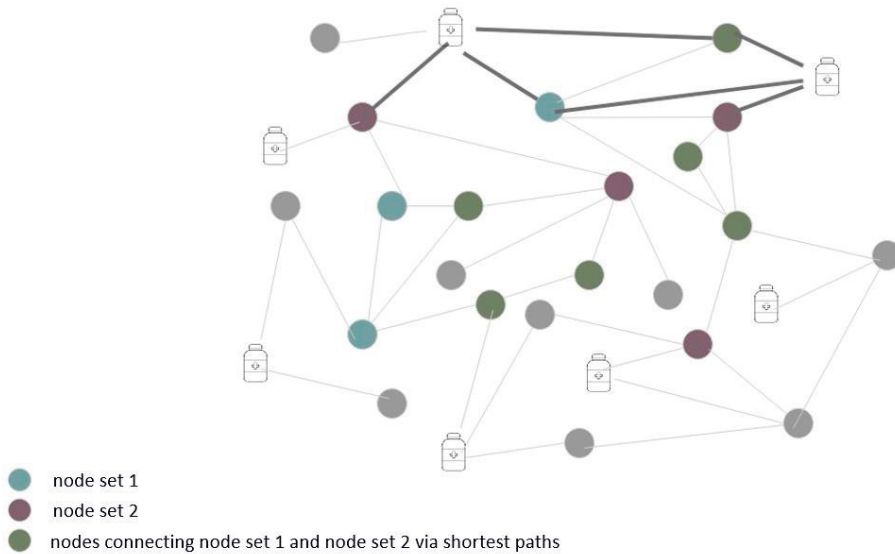
### 8.1.3 Linking Between Independent Data Sets and Data Points

Additionally, the UKS makes it possible to infer links between independent data sets, that are not visible in the individual data, but only become visible when integrated across multiple data types. Saarimäki et al. (Saarimäki, Fratello, et al., 2023; Saarimäki, Morikka, et al., 2023) linked AOP KEs to known gene sets, such as pathways, phenotypes or Gene Ontology terms, to infer possible genes linked to specific KE terms. The AOPs as well as the KE annotations are integrated into the UKS as showcased in Figure 20. Through path analysis in the UKS these KE terms can be associated to individual genes via the KEY\_EVENT-[MAPPED\_TO]-TERM-[ASSOCIATED\_WITH]-GENE meta path, which allows to create a gene level annotation for KEs.



**Figure 20** How meta-paths (bold) in the UKS are used to infer genes associated with key events. Figure adapted from (Saarimäki, Morikka, et al., 2023).

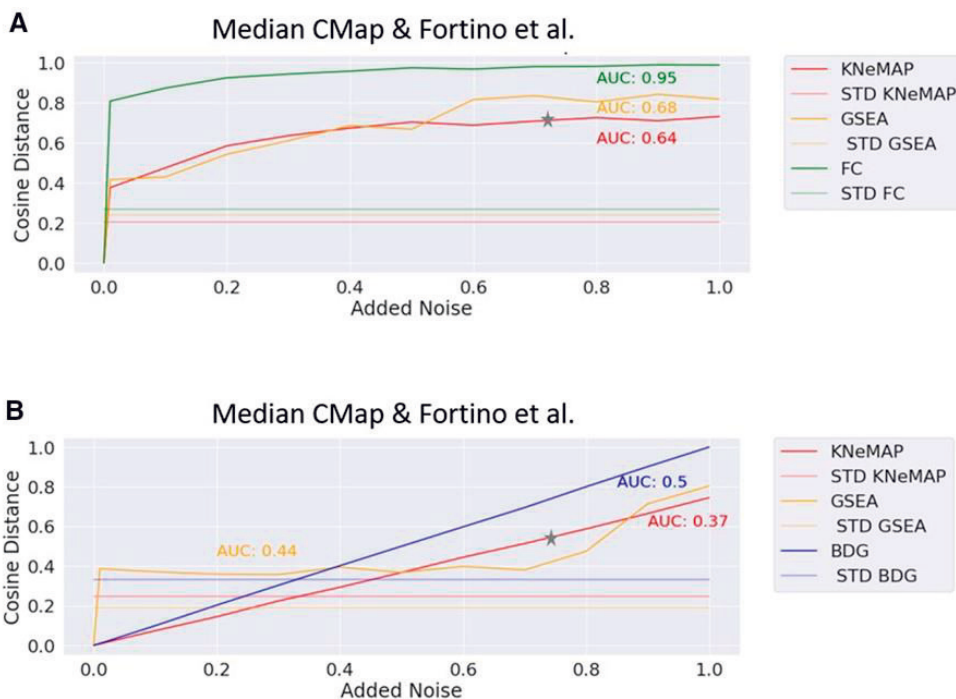
In publication I possible drug repositioning candidates for COVID-19 are identified by retrieving drugs from the UKS, known to target genes linked to all measured and inferred stages of a SARS-CoV-2 infection (Figure 21). This suggested drugs, targeting the opioid receptor and the coagulation cascade, as well as HDAC and proteasome inhibitors as possible drug candidates.



**Figure 21** UKS sub-network, containing gene (product) nodes, their interactions as well as drug nodes and drug gene target edges. Paths between genes, associated with different stages of a SARS-CoV-2 infection (red, blue, green), targeted by drugs are indicated by bold edges.

## 8.2 Robust Prior Information Can Reduce Noise in Gene Expression Data

KNeMAP (publication III), maps differential gene information to a robust multidimensional prior network (section 8.1.2). This method reduces the impact individual genes have on the comparative analysis between exposures or between datasets and therefore reduces the overall impact noisy data points can have across the analysis in comparison to other methods (Figure 22).



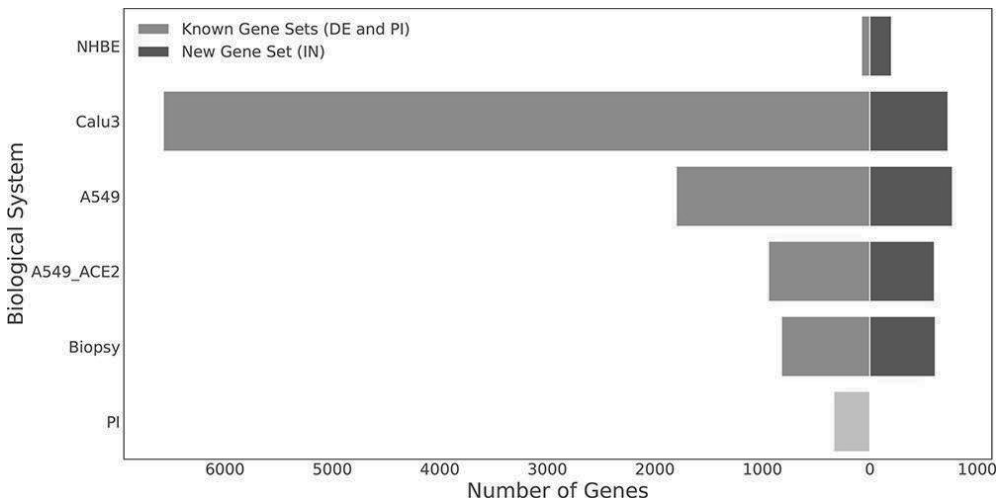
**Figure 22** Showcases the impact artificial added noise to gene expression data has on the stability of MOA vectors, computed with KNeMAP, gene deregulation analysis, fold changes and GSEA. KNeMAP's lowest AUC score indicates the least divergence from the baseline. Figure taken from publication III.

### 8.3 Identifying Non-Measured Relevant Genes

During the early stages of the COVID-19 pandemic, not much information about the MOA of COVID-19 or its long-term effects have been known, and relevant datasets were sparse. Gene expression data only measures a point in time and differential gene expression analysis only allows to identify genes, that change statistical significantly between measured time points. Molecular processes taking place in between these measured time points may stay undiscovered. In publication I two primary data types are available, one comprising genes, known to directly interact with SARS-CoV2 (Gordon et al., 2020) and transcriptomics data for multiple biological systems (Blanco-Melo et al., 2020), on which differentially expressed genes are computed (Ritchie et al., 2015). Based on the robust PPI network retrieved from the UKS (s. section 8.1.1), in combination with network analytical methods (chapter 4), it is possible to identify a set of genes, linking the



measurable sets of genes. This methodology identifies additional genes for each of the five available experimental gene expression datasets (Blanco-Melo et al., 2020), which significantly increases the set of genes possibly associated to a SARS-CoV2 infection (Figure 23).



**Figure 23** Shows the number of genes known to be associated to COVID-19 (at the time of publication) and how it can be increased with network analytics (set of IN genes). Figure taken from publication I.

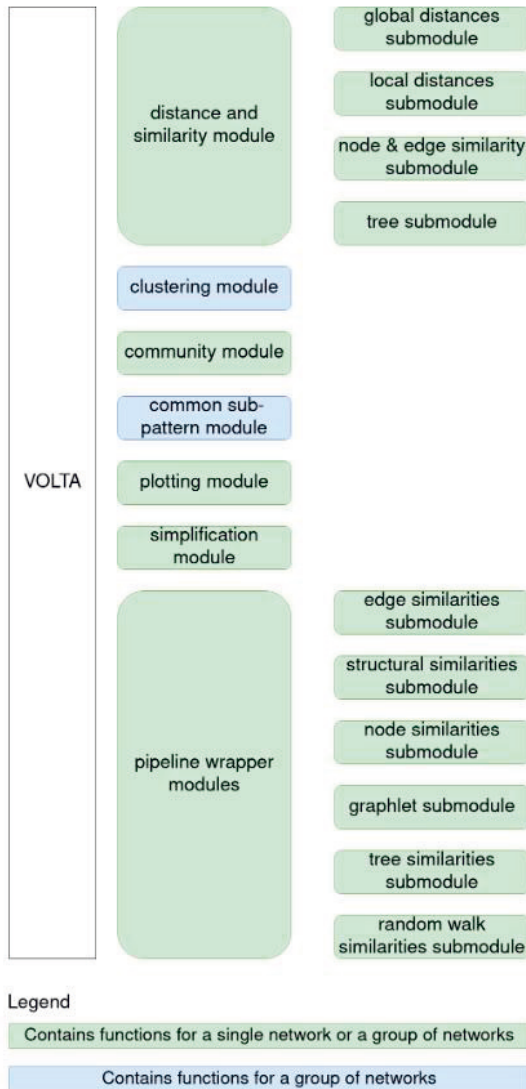
Analysing the biological processes of these identified genes, provided further insight into the pathogenesis of COVID-19, mostly suggesting an impact on vascular function of COVID-19, which is not significantly visible based on either the virus gene targets, or the measurable differential expressed genes.

## 8.4 A Comprehensive Network Analysis and Comparison Library

VOLTA (publication II) is a comprehensive Python package for network analysis, with a focus on multi gene co-expression network analysis. It is tuned to compare networks with each other, to cluster networks based on different similarity measures as well as to extract similarity patterns between networks. However, the individual methods and functions can be applied to a wide variety of networks due to its open-source and function focused implementation.

## 8.4.1 Modules

The VOLTA API consists out of seven main modules, of which some contain additional sub-modules. In total there are 158 exposed functions distributed over seven main modules and 10 submodules, as showcased in Figure 24.



**Figure 24** Showcases the modules and sub-modules implemented in VOLTA. Figure taken from publication II.

## Network Similarity and Distance Module

The *network similarity and distance* module combines four sub-modules, focusing on different aspects of how networks can be described and similarities between them measured. The *global distance* sub-module contains topological network measures, which describe a network in its whole, instead of focusing on individual genes. The *local distance* sub-module focuses on exploring individual nodes and their neighbourhoods as well as comparing them between networks. The *node and edge similarity* sub-module provides different distance metrics, which can be applied to nodes, edges and to compute similarity scores between networks. The *tree* sub-module converts the network structure into a binary tree and allows the application of different binary tree metrics to describe the network. This can be especially useful for networks with a loop-based structure (Giarratano et al., 2020). In total there are 53 exposed functions across four sub-modules.

## Network Clustering Module

The *network clustering* module contains different clustering algorithms, which take as input the distance matrices and metrics, computed in the *network similarity and distance* module. In addition, it also provides a multi-objective function to tune clustering parameters, based on the user-set objectives.

## Community Module

In the *community* module different community detection algorithms, a consensus method (Tandon et al., 2019) as well as different community evaluation functions are provided. The module contains 49 exposed functions.

## Identification of Common Sub-patterns Module

This module takes a set of networks or multiple groups of networks as input and extracts common or statistical overrepresented sub-graphs from them. In addition, a method to compute a consensus community partitioning across a set of networks is provided. The module is constructed out of seven exposed functions.

## Network Simplification Module

The module contains different functions to simplify complex networks by node or edge removal. It provides different methods on how to estimate weak links in the network as well as functions for edge weight adjustment. The module exposes seven functions in total.

## Plotting Module

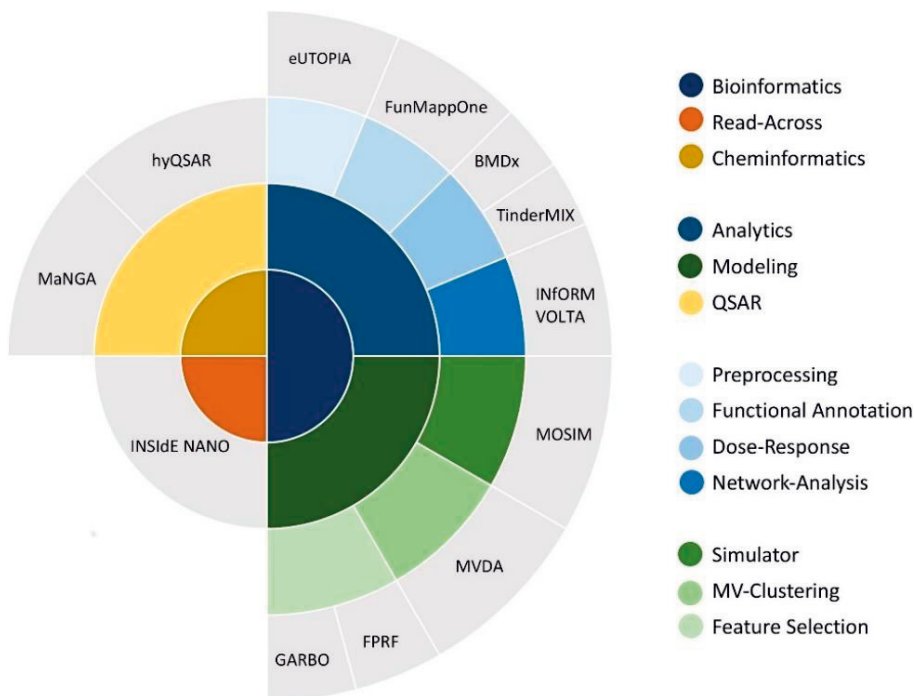
This module implements multiple functions which can plot the results of the other modules and contains a total of seven exposed functions.

## Pipeline Wrappers Module and Analysis Pipelines

The module provides wrappers to call a group of functions, that lead to a specific intermediate analysis step as well as complete example pipelines, in the form of Jupyter Notebooks, which allow for simple modification and adjustment by the user. The module consists out of 22 exposed functions, three analysis pipelines and three additional tutorial Jupyter Notebooks.

### 8.4.2 Integration into a Toxicogenomic-Analysis Pipeline

Publication II VOLTA, as a network analysis Python package, especially tuned for the comparison and analysis of multiple gene co-expression networks can be used to analyse and interpreted toxicogenomic data in combination with other toxicogenomic software. The NEXCAST software suite (Serra, Saarimäki, et al., 2022), showcases how VOLTA can be integrated and combined with different toxicogenomic software into a comprehensive software suite. NEXTCAS<sup>T</sup> offers diverse software to analyse and model toxicogenomic data, that can be combined in diverse manner to suit available data and analysis goals. The complete software suite is displayed in Figure 25.



**Figure 25** The NEXCAST software suite, combining multiple software needed for comprehensive toxicogenomic analysis, including VOLTA as a network analysis and comparison module. Figure taken from (Serra, Saarimäki, et al., 2022).

## 8.5 Differential Analysis of Co-Expression Networks

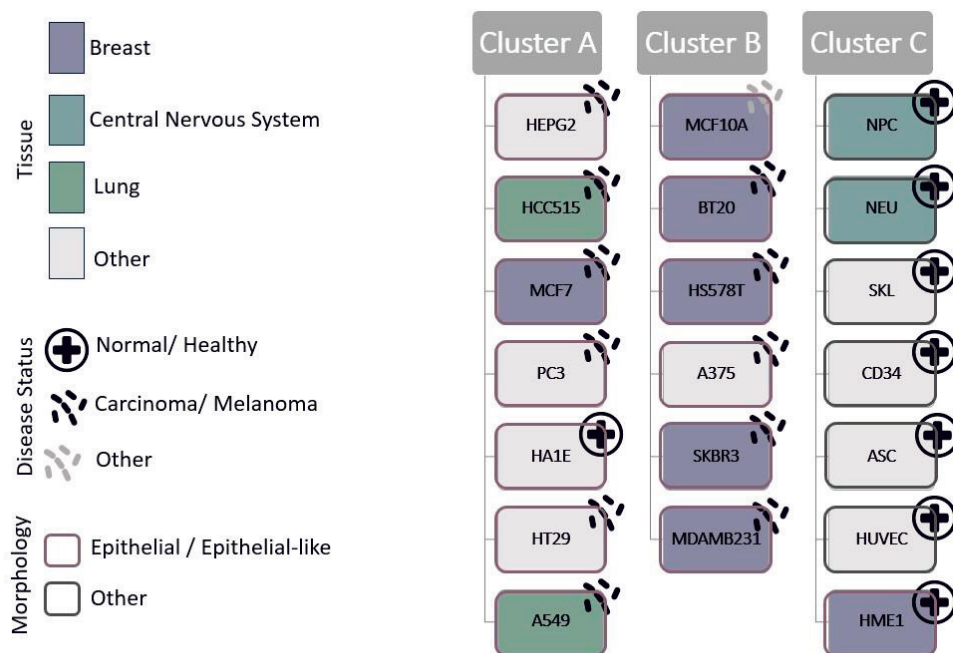
The MOA of a condition versus a control can be described in a co-expression gene network, by characterizing which genes change in their network influence with respect to the whole system. In publication II, the efficiency of such a differential centrality analysis is shown, by showcasing how this method can correctly identify genes related to the known MOA of specific drugs (the networks and algorithm parameters are described in section 7.3.1). The effect of dasatinib and mitoxantrone on A549 cell lines are compared and it can be observed that *OXA1L*, *DNAJC15* and *YME1L1* play a role in the mitoxantrone network, which are linked to mitochondrial protein metabolism impairment (Rossato et al., 2014).

In Federico et al. (Federico, Pavel, et al., 2022) the methodology is applied to identify genes that play a possible role in the outbreak of psoriatic lesions and further shows that this methodology can identify possible genes of interested, that are not visible

with traditional differential gene expression analysis methodologies (Federico et al., 2020).

## 8.6 Characterizing the Impact of Exposure System on a Compound's MOA

To showcase the importance of considering the exposure system when comparing the MOA of different compounds, it is investigated how different exposures across biological systems cluster. For 20 different biological systems, gene co-expression networks are created and grouped by means of different network metrics and clustering algorithms (as described in section 7.3.2). Publication II VOLTA identifies three different cell line clusters, which have different responses when treated with dasatinib. The cell lines within each cluster can be described by their tissue of origin as well as disease status as described in Figure 26. These results indicate that it is necessary to take the differences between exposure systems into account when characterizing and comparing MOAs of compounds.



**Figure 26** The three MOA clusters detected with VOLTA across 20 different cell lines treated with dasatinib. Cluster B contains mainly cell lines derived from breast tissue, cluster C only cell lines derived from healthy tissue and cluster A contains mostly epithelial (like) carcinoma cells.

### 8.6.1 Identifying Compounds with Similar MOA Across Biological Systems and Across Data Sets

To identify compounds, with a similar MOA when exposed on the same biological system but which can differ when exposed on different biological systems, 676 drugs across three biological systems are analysed with KNeMAP (publication III) and clustered with VOLTA (publication II), as described in section 7.4. This identifies, a set of 38 compounds, which show a similar MOA on MCF7, PC3 as well as HL60 cell lines, by grouping into the same cluster when each biological system is analysed independently, and the clusters are compared. However, when clustering the MOA of the 38 compounds across all biological systems, the systems induce a different MOA, indicated by the grouping in exposure system rather than exposure compound. When characterizing the identified compounds, a prevalence of antimicrobial and antiarrhythmic agents as well as antipsychotics and hsp90 inhibitors is found. These compounds further can be linked to cancer treatment or are under investigation for cancer treatment (the three investigated exposure systems

are different cancer cell lines) by inducing a cytotoxic or cytostatic effect (Becker & Banik, 2014; Jafari et al., 2022; Kepp et al., 2012; Kingston, 2009; Majolo et al., 2019; Unsal-Beyge & Tuncbag, 2022; Weissenrieder et al., 2019; C.-H. Wu et al., 2016).

## 8.7 Linking Engineered Nanomaterials and Drugs based on their MOA

ENMs are more recently developed and therefore less described and understood compounds but show large potential usage in the industry as well as in medicine (Saarimäki et al., 2021). Identifying similarities between ENMs and already characterized compounds can therefore provide valuable insights into the MOA of ENMs and potential toxicity can be inferred. Publication III, KNeMAP, allows the comparison of exposure profiles across datasets by using a data independent multi-dimensional prior network (s. section 0). With the help of KNeMAP it is possible to identify drugs with a similar MOA in the CMap dataset (Lamb et al., 2006) to the ENM core materials used in the Fortino et al. dataset (Fortino et al., 2022), as listed in Table 2.



**Table 2** ENM core materials and known drugs with a similar MOA, computed across datasets with publication III KNEMAP.

ENM Core Material	Drug	MOA
Copper oxide	Lycorine	Lycorine, is an acetylcholinesterase inhibitor (Kola et al., 2023), which plays a role in nerve impulse transmission. Copper oxide has been linked to affect acetylcholinesterase levels and is known to have a potential effect on the nervous system (Ganesan et al., 2016; Sezer Tuncsoy et al., 2019) .
Silver	Ciclopirox	Both silver and Ciclopirox have been used as antimicrobial treatments (L. Xu et al., 2020) .
Gold	Oxyphenbutazone	Gold compounds and gold nanoparticles have been used as treatment for inflammatory diseases, such as rheumatoid arthritis (Hornos Carneiro & Barbosa, 2016), while Oxyphenbutazone is not in use anymore, due to potential adverse outcomes, it was in use as an anti-inflammatory agent to reduce symptoms of arthritis (National Center for Biotechnology Information, n.d.).
Titanium dioxide	Nortriptyline	Both compounds are known to influence neurotransmitters, where nortriptyline is used as an antidepressant (Merwar et al., 2023; Naima et al., 2021) .
Multi-Walled Carbon Nanotubes	Thioridazine	Thioridazine has been taken off the market due to its potential to cause liver injury (“Thioridazine,” 2012), similar it is known that multi-walled carbon nanotubes are hepatotoxic (Z. Ji et al., 2009; Sun et al., 2021) .

## 9 DISCUSSION

Alternative methods for chemical safety assessment, drug repositioning and development as well as the personalized medicine concept rely on or need to be supported by computational methodologies, prior knowledge and available data (Pavel, Saarimäki, et al., 2022; Serra et al., 2020; Serra, Fratello, et al., 2022). Data is one of the most valuable currencies of our time and while many industries are at the top of mining, analysing and learning from large scales of data, the life science domain is only at the beginning of it.

Making efficient use of the data available across the different life science sub-fields can provide in depth understandings of underlying processes (Federico et al., 2020; Federico, Pavel, et al., 2022; S. Gao et al., 2021; P. Kinaret et al., 2017; Saarimäki et al., 2020; Serra et al., 2019), improve patient care (Cirillo & Valencia, 2019; D’Onofrio et al., 2022; Dash et al., 2019; Gu et al., 2021; Kovačević et al., 2020; Kruger et al., 2019; Kurt et al., 2008; Rajkomar et al., 2018; Taramasco et al., 2019; Zame et al., 2020) or the safety of materials, such as industrial chemicals, drugs and engineered nanomaterials for humans and the whole environment (J. Chen et al., 2021; Mouchlis et al., 2021; Sharifi et al., 2021; Sharma et al., 2023; X. Tong et al., 2021).

However, while a lot of data is currently available and is constantly produced, there are many unique challenges associated with Big Data integration in the life sciences, due to the differences in standards, technologies as well as the lack of a data centred view during data creation (Leonelli, 2019; Marx, 2013; Pavel, Saarimäki, et al., 2022). Additionally, the data is of high complexity, ever evolving and comprises many different data types, standards, and formats, which makes traditional data models and integration methodologies unsuitable when used on a large amount of these data.

Here the Unified Knowledge Space is presented, a Knowledge Graph framework comprising a graph data model, connected to a file storage, uniquely adjusted to life-sciences data, with a focus on chemical and drug relevant data points. I have identified different data types available as well as more than 80 independent data sources that can be processed and integrated with a mix of computational and manual data curation and entity identification methodologies. For each data type I

have identified a reference identification system, which is comprehensive and computational available to allow supervised automatic entity identification. For data types where such a reference system is not available, I have showcased the manual curation of an entity vocabulary and how manual curation in combination with NLP based software (Di Lieto et al., 2023) can be used to maintain the vocabulary as well as to perform automated entity identification. Based on these approaches, together with the developed data model, it was possible to create a life science Big Data knowledge based. This knowledge base can be applied to many different problems across the life sciences as well as can be adjusted and expanded into additional sub-fields and research areas through the integration of additional data types and sources, as well as is flexible enough to evolve with the data creation technologies. The UKS is to my knowledge the largest life-sciences knowledge graph data model created to date (Abdelaziz et al., 2017; Al-Saleem et al., 2021; Z. Chen et al., 2022; Mohamed et al., 2019; R. Zhang et al., 2021), by including over three billion data points.

The data integration task was especially challenging for data/ entity types for which no standardized reference systems exist or have been used by the publisher of the data. For example, phenotypic data are often reported as strings, which when needed to be mapped between data sources can either be done based on NLP methods or due to manual curation. NLP based approaches have the advantage of automatic this task, however, do not have the professional understanding of identifying if two phenotypes should be considered the same or maybe be a subgroup of each other, while for manual curation a lot of field specific knowledge is needed. For example, *Diabetes Type 1* and *Diabetes Type 2* will receive a high string-matching score based on NLP methodologies however based on the granularity of the data system may be considered as different diseases. Similar issues arise with other entities often reported as strings, such as cell lines, cell types or tissues. Drugs and chemicals if reported under their government approved, commercial name or based on their chemical structure can mostly be identified, however custom compounds that are not widespread available will likely not appear in any other data source in the future. However, if structural information is available, they can be matched to similar compounds in the future. The mapping between genes and gene product identification systems can be mostly automated due to the widespread use of common identification systems, such as Entrez (Maglott et al., 2011) and Ensembl (Cunningham et al., 2022) for which while no 1-1 mapping exists, many public mapping APIs, such as mygene (C. Wu et al., 2013), exist. Further it has been widely accepted to report genes and gene products on these identification systems and the less consistent

identification by gene symbol is mostly used in-text and for highly described genes or gene products. Lastly the least challenging data types to integrated are these based on widely accepted source systems, such as Gene Ontology (The Gene Ontology Consortium, 2021) terms, pathways (Jassal et al., 2020; Kanehisa et al., 2017; Martens et al., 2021) or AOPs, which when reported based on their source system IDs are easy to identify. However, if the less standardized and consistent names are used the mapping relies again on NLP based approaches and not all conflicts can be resolved. In conclusion the large-scale data integration challenges is dependent on the way entities are reported across data sources, and standardized IDs should be used when available.

The UKS contrasts with many other KGs available in the field (Abdelaziz et al., 2017; Al-Saleem et al., 2021; Chandak et al., 2023; Z. Chen et al., 2022; Che et al., 2021; Z. Gao et al., 2022; Jiajing Hu et al., 2021; Karim et al., 2019; Mohamed et al., 2020, 2019; Myklebust et al., n.d.; Nováček & Mohamed, 2020; Sosa et al., 2020; Meng Wang et al., 2021; Shudong Wang et al., 2022; F. Zhang et al., 2021; R. Zhang et al., 2021; Zheng et al., 2021; Y. Zhu et al., 2020) by not being created for a specific study, hypothesis or problem but rather a general framework on which a multitude of studies can be performed. This implies that data search, data adjustment or scope adjustment can be performed highly efficient on the UKS, since the individual datasets do not need to be curated or identified from different sources but can rather be retrieved in an already curated and unified manner from the UKS. In the example of meta-analysis, it has been shown that the most expensive step is the identification and combination of relevant data sets or studies and by using data management methods their cost can be reduced, while flexibility and speed can be improved (Tiddi et al., 2020).

Due to the combined vocabulary of the UKS, as well as its management of experimental datasets, it allows the fast search and retrieval of data for a certain condition without the need of curating the data to use the same identifiers or be in the same format and therefore having the potential to significantly improve speed and reduce cost of a diverse set of data-dependent (toxicology and pharmacology) analyses. In addition, the UKS allows to retrieve additional information and data-points about entities or processes under investigation without the need to first identify suitable external resources and likely needing to convert the current used identifiers to the identifiers used in the external system. Such information can be used to enrich analysis results or to understand the implications of multiple or individual data points/ entities for the condition under investigation. Additionally, the output format can be retrieved in the same format as previously retrieved data points on which the analysis may have been performed. Publications I and III as well

as Federico et al. (Federico, Fratello, et al., 2022) and Saarimäki et al. (Saarimäki, Fratello, et al., 2023; Saarimäki, Morikka, et al., 2023) made use of the unified nature of the UKS, by retrieving different datasets, in the same format, from the UKS in the form of gene-gene similarity networks and a robust PPI network or enriching entities with further information, such as drugs targeting a set of gene(s) (products) or linking AOP KEs to gene (products).

While in these cases the UKS allowed fast data retrieval and knowledge lookup it also restricts the research to the data points and sources integrated (at the point of analysis) in the UKS and in result introduces a data availability bias into the analysis. However, selecting and combining data sources for each new analysis when needed, may lead to a different selection of data sources, but likely also to a smaller selection of sources due to the manual effort needed in the curation step and in result propagating the biases existing in the selected data source(s). For example, Tang et al. (Tang et al., 2019), Guo et al. (Guo et al., 2022), Prashanth et al. (Prashanth et al., 2021) and Tu et al. (Tu et al., 2022) supported, enriched or performed their analysis with/ on PPI data collected from only one data source.

While the general nature of the UKS has many advantages, it could reduce predictive power of individual analysis if the sub-graphs used for the analysis are not selected problem specific. It has been shown that KG predictive performance improves with the problem specificity of the KG (Ratajczak et al., 2022) However, the UKS allows the retrieval of sub-graphs, which allows to create problem specific networks, but this requires the user to identify the most relevant data-points for their study in the UKS.

Creating and maintaining such a large-scale general knowledge space requires a lot of manual effort during the data engineering process. Especially in a schema free framework, such as the UKS, the database administrator is responsible to keep the knowledge space as clean and understandable as possible, as well as to maintain transparency of the individual data points. Additionally, such a large knowledge space requires large amounts of computational power with respect to disk space and memory, which is required to stored and load the UKS, while large amounts of computational power, such as central processing units (CPU) and possibly graphics processing units (GPU) are needed to perform large scale analyses on the UKS. Scaling computational power can quickly become expensive and the resources required for such a framework are not always available.

With the continuous production of data across the life sciences as well as with the updating of existing data sets, the UKS needs to be constantly maintained, adapted and new data sets need to be integrated. Next to the hardware needed, this also requires personnel dedicated to the maintenance of such a general knowledge base. However, many projects are funded for a limited time only which can lead to such general framework not being able to be continued or maintained for a long time. While there have been automated approaches suggested for the creation of KGs (Lobentanzer et al., 2023), they often require a pre-defined schema, which severely limits the adaptability of a KG to different data sources and sets as well as neglect the entity identification and mapping challenge by either only accepting a set of identification systems or separating namespaces.

While developing the UKS, I particularly emphasized entity mapping, which requires a combination of automated (NLP based) methods together with manual efforts. This effort contrasts with other proposed approaches (Lobentanzer et al., 2023), where emphasis is put on automation rather than combining namespaces. Keeping namespaces separate however severely limits the linked power of KGs, by only allowing to link between knowledge coming from data sources which already use the same namespace. This approach allows for duplicate entities, when they make use of different namespaces, which can complicate the data retrieval process as well as the knowledge inference process, since the user either needs to perform entity mapping at a later stage or select a namespace for their analysis. This limits the analysis to data linked to that namespace only, meaning other information linked to a different namespace are not included in the analysis. Keeping namespaces separated can further propagate or introduce biases into the analysis, since the available data is limited as well as are namespaces often common to specific research areas or regions, which will emphasise their native data biases in the performed analysis. For example, European data points will likely be linked to Ensembl Gene identifiers (Cunningham et al., 2022), while (US) American research outputs are likely linked to Entrez gene identifiers (Maglott et al., 2011).

Further a general framework implies, that all the decisions taken during the data integration and curation stage will be propagated towards all analyses performed on the knowledge space. This can imply that possible errors introduced during the curation stage may be propagated into many analyses before they are identified. However, collecting and curating data for each analysis individually increases the risk of introducing errors, since the steps are performed multiple times (for each analysis) instead of a single time only. Further the Big Data approach in the UKS, allows to possibly identify erroneous data points (s. Section 5.1.1 (Assessing Data Point

Quality via Network Topology)), which may be introduced during the data creation or data curation step.

The UKS can be seen as a proof of concept of the possibility to create highly flexible, multi-billion data point Big Data models, which can easily be expanded to include additional data from other life science sub-domains. Currently the UKS is chemical centred, due to the nature of the applied case studies, but with the addition of new datasets the focus can be broadened or shifted entirely without the need to design a new data model or infrastructure. The here presented case studies only focus on small aspects of the UKS to showcase its possibilities in combination with network analytical methodologies, and the suitability of the data model for multiple different application scenarios. However, the UKS has further potential, especially as a predictive engine and underlying data source for deep learning models based on multiple millions to billions of data points (Nickel et al., 2016), which will be explored in the future.

The work presented in this thesis, showcases how with a combination of computational approaches, manual efforts as well as flexibility, large scale multidimensional knowledge bases can be created in the life sciences. While the life sciences are made up of different sub-fields, that use different methodologies and standards, they are however not stand-alone fields but rather contribute to each other. Therefore, to make use of the available data to its full potential they need to be integrated, analysed and learned from in a combined manner. This thesis showcases how Knowledge Graphs and graph data models can be used for this task, while in addition to their highly flexible nature, allows the interpretation and analysis of data points as part of a connected system rather than on their own. The paradigm that entities, processes, events and their outcomes can only be understood as part of the whole system is a concept that has become more and more prevalent in different areas of the life sciences (chapter 4), which the UKS, with its network-based data model, together with VOLTAs network analytical methodologies supports natively. For example in toxicology AOP networks are used to understand what mechanisms take part between observable triggers and (phenotypic) outcomes (Ankley et al., 2010; Knapen et al., 2018), in systems biology different interaction and regulation networks are used to understand molecular processes in depth (Albert, 2007; Gosak et al., 2018; Yan et al., 2018) and in the neurosciences neuronal networks are used to understand the brain (Bullmore & Sporns, 2009; Meng & Xiang, 2018), while in ethology interaction networks between species and individuals are studied (Gosztolai & Ramdya, 2022; Makagon et al., 2012; Wey et al., 2008).

VOLTA aims in comparison to other comparable network analytical tools to be Graphical User Interface free (Kuntal et al., 2016; Marwah et al., 2018; Proost & Mutwil, 2018) and rather expose all main functionalities to not restrict applicability, adjustability or importable data. While providing pre-defined analysis pipelines for gene co-expression network analysis VOLTA also provides many analysis and machine learning methodologies, such as clustering, that can be used on (biological) networks but also in different application scenarios, such has been done in publication III, where VOLTA is used to cluster gene expression vectors in addition to perform community detection on a gene gene similarity network extracted from the UKS. This allows VOLTA to be used in a wide range of applications and datasets and not be limited by implementation specific factors, such as pipeline restrictions or missing function exposure. Further VOLTA is one of the few (biological) network analysis packages, which supports the grouping of multiple networks and the computation of similarities between networks (Csardi & Nepusz, 2006; Hagberg et al., 2008; Marwah et al., 2018). Gene co-expression networks are often used to identify the MOA of an exposure or phenotype (Federico, Pavel, et al., 2022; P. Kinaret et al., 2017; Koenig et al., 2021; Song et al., 2019), however identifying groups of similar acting compounds or similarities in gene co-expression profiles across compounds and phenotypes can for example help in identifying compounds (or sub-structures) which may lead to similar MOAs or identifying drugs that show similar/ opposite characteristics to phenotypes which can be used during drug repositioning studies or the chemical safety assessment process. In publication II this feature has been used to showcase that cell line/ tissue of origin has a strong effect on the behaviour of a drug, and therefore biological system information must be considered when interpreting transcriptomic data.

Across different case studies I have presented how different aspects of the UKS in combination with network analytical and toxicogenomic methodologies can be used to a) support and improve toxicogenomic analysis through robust data retrieval and b) how it can be used to infer new data points and knowledge. The main contribution of this thesis work is to showcase the possibility and applicability of diagonal Big Data integration in the life sciences to create a problem unspecific knowledge base, which can be achievable by combining different computational and manual curation, modelling and analysis methodologies, which contrasts with many currently available approaches.



## 10 SUMMARY AND CONCLUSION

Big Data integration and analytics have become highly popular over the last decade, especially in the IT industry. Large IT companies, such as Meta, Google, Amazon and Netflix are built on the idea of gaining insight into user behaviour through collecting and analysing user data on a large scale. Also in the life sciences, Big Data has gained interest in recent years (Z. Chen et al., 2022; Fröhlich et al., 2018; Gu et al., 2021; Pavel, Saarimäki, et al., 2022; Y. Wu et al., 2019; Zong et al., 2022).

However, due to the fractured nature of the field, the non-standardized data reporting and creation standard and in general the large diversity across sub-disciplines have made the large-scale usage of data across disciplines challenging (Leonelli, 2014, 2019; Pavel, Saarimäki, et al., 2022).

Network analytics have been widely applied in recent years, especially in systems biology, and have shown their potential to gain in-depth insights into molecular processes (Badkas et al., 2021; Guo et al., 2022; P. Kinaret et al., 2017; Y. Liu et al., 2019; Marwah et al., 2018; Pavel, Serra, et al., 2022; Serra et al., 2019). KGs have been applied on a small scale in different life science disciplines (Abdelaziz et al., 2017; Chandak et al., 2023; Y. Chen et al., 2021; Z. Chen et al., 2022; Z. Gao et al., 2022; Karim et al., 2019; Pavel, Saarimäki, et al., 2022) and together with network analytical methods have shown their potential to infer valuable data driven insights into the problem under investigation (Z. Gao et al., 2022; Karim et al., 2019; R. Zhang et al., 2021). However, due to the challenging nature of the data, complexity of the data integration and data modelling task, there have mostly only been small, problem specific KGs and data analysis task created to date (Pavel, Saarimäki, et al., 2022).

Here, I presented the Unified Knowledge Space, as a proof of concept, that large scale manual supported data integration across multiple life science sub-domains to model multiple billions of data points is possible, when a highly flexible but strictly maintained data model is used. I have further showcased how different sub-graphs of the UKS, in combination with network analytical concepts can be used to improve toxicogenomic data, analyse and compare it or infer additional knowledge about the problem under investigation. I have shown how the UKS can be used as a 1) integrated data source for data retrieval and entity mapping, 2) how its modelled data

can be analysed with network analytical methods and 3) how it can be used to infer facts that are either only visible through the path based data model or can be computed by applying a reasoning engine (for example in the application of the AOP KE gene annotation process or drug repositioning for COVID-19).

The full potential of the UKS has not been exploited yet, especially on its potential to infer new facts about the world under investigation. But due to the complexity, size and ever-growing nature of the UKS, these are continuously ongoing tasks. This further shows that the UKS and its applicability are not statics but ever evolving and I predict the here presented data model to be used in many further studies and to gain ever more detailed insights into complex biological processes and the interaction of compounds with living beings and the environment.

## REFERENCES

- Abdelaziz, I., Fokoue, A., Hassanzadeh, O., Zhang, P., & Sadoghi, M. (2017). Large-scale structural and textual similarity-based mining of knowledge graph to predict drug–drug interactions. *Web Semantics: Science, Services and Agents on the World Wide Web*, 0(0). <https://doi.org/10.1016/j.websem.2017.06.002>
- Aboul-Yazeed, R. S., El-Bialy, A., & Mohamed, A. S. A. (2017). Prediction of medical equipment failure rate: A case study. In A. E. Hassanien, K. Shaalan, T. Gaber, A. T. Azar, & M. F. Tolba (Eds.), *Proceedings of the international conference on advanced intelligent systems and informatics 2016* (Vol. 533, pp. 650–659). Springer International Publishing. [https://doi.org/10.1007/978-3-319-48308-5\\_62](https://doi.org/10.1007/978-3-319-48308-5_62)
- Admon, A. J., Donnelly, J. P., Casey, J. D., Janz, D. R., Russell, D. W., Joffe, A. M., Vonderhaar, D. J., Dischert, K. M., Stempek, S. B., Dargin, J. M., Rice, T. W., Iwashyna, T. J., & Semler, M. W. (2019). Emulating a novel clinical trial using existing observational data. predicting results of the prevent study. *Annals of the American Thoracic Society*, 16(8), 998–1007. <https://doi.org/10.1513/AnnalsATS.201903-241OC>
- Afzaal, H., Altaf, R., Ilyas, U., Zaman, S. U., Abbas Hamdani, S. D., Khan, S., Zafar, H., Babar, M. M., & Duan, Y. (2022). Virtual screening and drug repositioning of FDA-approved drugs from the ZINC database to identify the potential hTERT inhibitors. *Frontiers in Pharmacology*, 13, 1048691. <https://doi.org/10.3389/fphar.2022.1048691>

- Agarwal, V., Bell, G. W., Nam, J.-W., & Bartel, D. P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. *ELife*, 4. <https://doi.org/10.7554/eLife.05005>
- Al-Saleem, J., Granet, R., Ramakrishnan, S., Ciancetta, N. A., Saveson, C., Gessner, C., & Zhou, Q. (2021). Knowledge Graph-Based Approaches to Drug Repurposing for COVID-19. *Journal of Chemical Information and Modeling*, 61(8), 4058–4067. <https://doi.org/10.1021/acs.jcim.1c00642>
- Alanis-Lobato, G., Andrade-Navarro, M. A., & Schaefer, M. H. (2017). HIPPIE v2.0: enhancing meaningfulness and reliability of protein-protein interaction networks. *Nucleic Acids Research*, 45(D1), D408–D414. <https://doi.org/10.1093/nar/gkw985>
- Albert, R. (2007). Network inference, analysis, and modeling in systems biology. *The Plant Cell*, 19(11), 3327–3338. <https://doi.org/10.1105/tpc.107.054700>
- Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F., & Hamosh, A. (2015). OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Research*, 43(Database issue), D789–98. <https://doi.org/10.1093/nar/gku1205>
- Amberger, J. S., Bocchini, C. A., Scott, A. F., & Hamosh, A. (2019). OMIM.org: leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Research*, 47(D1), D1038–D1043. <https://doi.org/10.1093/nar/gky1151>
- Anglani, R., Creanza, T. M., Liuzzi, V. C., Piepoli, A., Panza, A., Andriulli, A., & Ancona, N. (2014). Loss of connectivity in cancer co-expression networks. *Plos One*, 9(1), e87075. <https://doi.org/10.1371/journal.pone.0087075>
- Ankley, G. T., Bennett, R. S., Erickson, R. J., Hoff, D. J., Hornung, M. W., Johnson, R. D., Mount, D. R., Nichols, J. W., Russom, C. L., Schmieder, P. K., Serrano, J. A., Tietge, J. E., & Villeneuve, D. L. (2010). Adverse outcome pathways: a conceptual

- framework to support ecotoxicology research and risk assessment. *Environmental Toxicology and Chemistry*, 29(3), 730–741. <https://doi.org/10.1002/etc.34>
- Argelaguet, R., Cuomo, A. S. E., Stegle, O., & Marioni, J. C. (2021). Computational principles and challenges in single-cell data integration. *Nature Biotechnology*, 39(10), 1202–1215. <https://doi.org/10.1038/s41587-021-00895-7>
- Arrell, D. K., & Terzic, A. (2010). Network systems biology for drug discovery. *Clinical Pharmacology and Therapeutics*, 88(1), 120–125. <https://doi.org/10.1038/clpt.2010.91>
- Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1), 25–29. <https://doi.org/10.1038/75556>
- Badkas, A., De Landsheer, S., & Sauter, T. (2021). Topological network measures for drug repositioning. *Briefings in Bioinformatics*, 22(4). <https://doi.org/10.1093/bib/bbaa357>
- Bai, Q., Liu, S., Tian, Y., Xu, T., Banegas-Luna, A. J., Pérez-Sánchez, H., Huang, J., Liu, H., & Yao, X. (2021). Application advances of deep learning methods for de novo drug design and molecular dynamics simulation. *Wiley Interdisciplinary Reviews: Computational Molecular Science*. <https://doi.org/10.1002/wcms.1581>
- Ball, T., Barber, C. G., Cayley, A., Chilton, M. L., Foster, R., Fowkes, A., Heghes, C., Hill, E., Hill, N., Kane, S., Macmillan, D. S., Myden, A., Newman, D., Polit, A., Stalford, S. A., & Vessey, J. D. (2021). Beyond adverse outcome pathways: making toxicity predictions from event networks, SAR models, data and knowledge. *Toxicology Research*, 10(1), 102–122. <https://doi.org/10.1093/toxres/tfaa099>

- Bandrowski, A., Brinkman, R., Brochhausen, M., Brush, M. H., Bug, B., Chibucos, M. C., Clancy, K., Courtot, M., Derom, D., Dumontier, M., Fan, L., Fostel, J., Fragoso, G., Gibson, F., Gonzalez-Beltran, A., Haendel, M. A., He, Y., Heiskanen, M., Hernandez-Boussard, T., ... Zheng, J. (2016). The ontology for biomedical investigations. *Plos One*, *11*(4), e0154556.  
<https://doi.org/10.1371/journal.pone.0154556>
- Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., Wilson, C. J., Lehár, J., Kryukov, G. V., Sonkin, D., Reddy, A., Liu, M., Murray, L., Berger, M. F., Monahan, J. E., Morais, P., Meltzer, J., Korejwa, A., Jané-Valbuena, J., ... Garraway, L. A. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, *483*(7391), 603–607.  
<https://doi.org/10.1038/nature11003>
- Battram, T., Yousefi, P., Crawford, G., Prince, C., Sheikhal Babaei, M., Sharp, G., Hatcher, C., Vega-Salas, M. J., Khodabakhsh, S., Whitehurst, O., Langdon, R., Mahoney, L., Elliott, H. R., Mancano, G., Lee, M. A., Watkins, S. H., Lay, A. C., Hemani, G., Gaunt, T. R., ... Suderman, M. (2022). The EWAS Catalog: a database of epigenome-wide association studies. *Wellcome Open Research*, *7*, 41.  
<https://doi.org/10.12688/wellcomeopenres.17598.2>
- Becker, F. F., & Banik, B. K. (2014). Polycyclic aromatic compounds as anticancer agents: synthesis and biological evaluation of methoxy dibenzofluorene derivatives. *Frontiers in Chemistry*, *2*, 55. <https://doi.org/10.3389/fchem.2014.00055>
- Bell, S. M., Angrish, M. M., Wood, C. E., & Edwards, S. W. (2016). Integrating publicly available data to generate computationally predicted adverse outcome pathways for fatty liver. *Toxicological Sciences*, *150*(2), 510–520.  
<https://doi.org/10.1093/toxsci/kfw017>

- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
- Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit* (1st ed., p. 504). O'Reilly Media.
- Blanco-Melo, D., Nilsson-Payant, B. E., Liu, W.-C., Uhl, S., Hoagland, D., Møller, R., Jordan, T. X., Oishi, K., Panis, M., Sachs, D., Wang, T. T., Schwartz, R. E., Lim, J. K., Albrecht, R. A., & tenOever, B. R. (2020). Imbalanced Host Response to SARS-CoV-2 Drives Development of COVID-19. *Cell*, *181*(5), 1036-1045.e9. <https://doi.org/10.1016/j.cell.2020.04.026>
- Boyle, E. I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J. M., & Sherlock, G. (2004). GO:TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, *20*(18), 3710–3715. <https://doi.org/10.1093/bioinformatics/bth456>
- Breuer, K., Foroushani, A. K., Laird, M. R., Chen, C., Sribnaia, A., Lo, R., Winsor, G. L., Hancock, R. E. W., Brinkman, F. S. L., & Lynn, D. J. (2013). InnateDB: systems biology of innate immunity and beyond--recent updates and continuing curation. *Nucleic Acids Research*, *41*(Database issue), D1228-33. <https://doi.org/10.1093/nar/gks1147>
- Bullmore, E., & Sporns, O. (2009). Complex brain networks: Graph theoretical analysis of structural and functional systems. *Nature Reviews. Neuroscience*, *10*(3), 186–198. <https://doi.org/10.1038/nrn2575>

- Butte, A. J., & Kohane, I. S. (2000). Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pacific Symposium on Biocomputing*, 418–429. [https://doi.org/10.1142/9789814447331\\_0040](https://doi.org/10.1142/9789814447331_0040)
- Castro-Mondragon, J. A., Riudavets-Puig, R., Rauluseviciute, I., Lemma, R. B., Turchi, L., Blanc-Mathieu, R., Lucas, J., Boddie, P., Khan, A., Manosalva Pérez, N., Fornes, O., Leung, T. Y., Aguirre, A., Hammal, F., Schmelter, D., Baranasic, D., Ballester, B., Sandelin, A., Lenhard, B., ... Mathelier, A. (2022). JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 50(D1), D165–D173. <https://doi.org/10.1093/nar/gkab1113>
- Ceschia, S., & Schaerf, A. (2011). Local search and lower bounds for the patient admission scheduling problem. *Computers & Operations Research*, 38(10), 1452–1463. <https://doi.org/10.1016/j.cor.2011.01.007>
- Chandak, P., Huang, K., & Zitnik, M. (2023). Building a knowledge graph to enable precision medicine. *Scientific Data*, 10(1), 67. <https://doi.org/10.1038/s41597-023-01960-3>
- Chang, Y., Park, H., Yang, H.-J., Lee, S., Lee, K.-Y., Kim, T. S., Jung, J., & Shin, J.-M. (2018). Cancer Drug Response Profile scan (CDRscan): A Deep Learning Model That Predicts Drug Effectiveness from Cancer Genomic Signature. *Scientific Reports*, 8(1), 8857. <https://doi.org/10.1038/s41598-018-27214-6>
- Chatr-aryamontri, A., Ceol, A., Palazzi, L. M., Nardelli, G., Schneider, M. V., Castagnoli, L., & Cesareni, G. (2007). MINT: the Molecular INTeraction database. *Nucleic Acids Research*, 35(Database issue), D572-4. <https://doi.org/10.1093/nar/gkl950>



- Chen, J., Si, Y.-W., Un, C.-W., & Siu, S. W. I. (2021). Chemical toxicity prediction based on semi-supervised learning and graph convolutional neural network. *Journal of Cheminformatics*, *13*(1), 93. <https://doi.org/10.1186/s13321-021-00570-8>
- Chen, Y., Ma, T., Yang, X., Wang, J., Song, B., & Zeng, X. (2021). MUFFIN: multi-scale feature fusion for drug-drug interaction prediction. *Bioinformatics*, *37*(17), 2651–2658. <https://doi.org/10.1093/bioinformatics/btab169>
- Chen, Z., Peng, B., Ioannidis, V. N., Li, M., Karypis, G., & Ning, X. (2022). A knowledge graph of clinical trials ([Formula: see text]). *Scientific Reports*, *12*(1), 4724. <https://doi.org/10.1038/s41598-022-08454-z>
- Che, M., Yao, K., Che, C., Cao, Z., & Kong, F. (2021). Knowledge-Graph-Based Drug Repositioning against COVID-19 by Graph Convolutional Network with Attention Mechanism. *Future Internet*, *13*(1), 13. <https://doi.org/10.3390/fi13010013>
- Cheng, F., Desai, R. J., Handy, D. E., Wang, R., Schneeweiss, S., Barabási, A.-L., & Loscalzo, J. (2018). Network-based approach to prediction and population-based validation of in silico drug repurposing. *Nature Communications*, *9*(1), 2691. <https://doi.org/10.1038/s41467-018-05116-5>
- Cheng, F., Kovács, I. A., & Barabási, A.-L. (2019). Network-based prediction of drug combinations. *Nature Communications*, *10*(1), 1197. <https://doi.org/10.1038/s41467-019-09186-x>
- Chirico, N., Sangion, A., Gramatica, P., Bertato, L., Casartelli, I., & Papa, E. (2021). QSARINS-Chem standalone version: A new platform-independent software to profile chemicals for physico-chemical properties, fate, and toxicity. *Journal of Computational Chemistry*, *42*(20), 1452–1460. <https://doi.org/10.1002/jcc.26551>

- Christensen, B., & Nielsen, J. (2000). Metabolic Network Analysis. In B. Sonnleitner (Ed.), *Bioanalysis and biosensors for bioprocess monitoring* (Vol. 66, pp. 209–231). Springer Berlin Heidelberg. [https://doi.org/10.1007/3-540-48773-5\\_7](https://doi.org/10.1007/3-540-48773-5_7)
- Cirillo, D., & Valencia, A. (2019). Big data analytics for personalized medicine. *Current Opinion in Biotechnology*, 58, 161–167. <https://doi.org/10.1016/j.copbio.2019.03.004>
- Coloma, P. M., Schuemie, M. J., Trifirò, G., Gini, R., Herings, R., Hippisley-Cox, J., Mazzaglia, G., Giaquinto, C., Corrao, G., Pedersen, L., van der Lei, J., Sturkenboom, M., & EU-ADR Consortium. (2011). Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project. *Pharmacoepidemiology and Drug Safety*, 20(1), 1–11. <https://doi.org/10.1002/pds.2053>
- Cozzoli, N., Salvatore, F. P., Faccilongo, N., & Milone, M. (2022). How can big data analytics be used for healthcare organization management? Literary framework and future research from a systematic review. *BMC Health Services Research*, 22(1), 809. <https://doi.org/10.1186/s12913-022-08167-z>
- Csabai, L., Fazekas, D., Kadlecsek, T., Szalay-Bekő, M., Bohár, B., Madgwick, M., Módos, D., Ölbei, M., Gul, L., Sudhakar, P., Kubisch, J., Oyeyemi, O. J., Liska, O., Ari, E., Hotzi, B., Billes, V. A., Molnár, E., Földvári-Nagy, L., Csályi, K., ... Korcsmáros, T. (2022). Signalink3: a multi-layered resource to uncover tissue-specific signaling networks. *Nucleic Acids Research*, 50(D1), D701–D709. <https://doi.org/10.1093/nar/gkab909>
- Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695(5), 1–9.
- Cunningham, F., Allen, J. E., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Austine-Orimoloye, O., Azov, A. G., Barnes, I., Bennett, R., Berry, A., Bhai, J., Bignell, A., Billis, K., Boddu, S., Brooks, L., Charkhchi, M., Cummins, C., Da Rin

- Fioretto, L., ... Flicek, P. (2022). Ensembl 2022. *Nucleic Acids Research*, 50(D1), D988–D995. <https://doi.org/10.1093/nar/gkab1049>
- D’Onofrio, A., Solimene, F., Calò, L., Calvi, V., Viscusi, M., Melissano, D., Russo, V., Rapacciuolo, A., Campana, A., Caravati, F., Bonfanti, P., Zanotto, G., Gronda, E., Vado, A., Calzolari, V., Botto, G. L., Zecchin, M., Bontempi, L., Giacomelli, D., ... Padeletti, L. (2022). Combining home monitoring temporal trends from implanted defibrillators and baseline patient risk profile to predict heart failure hospitalizations: results from the SELENE HF study. *Europace*, 24(2), 234–244. <https://doi.org/10.1093/europace/euab170>
- Dash, S., Shakyawar, S. K., Sharma, M., & Kaushik, S. (2019). Big data in healthcare: management, analysis and future prospects. *Journal of Big Data*, 6(1), 54. <https://doi.org/10.1186/s40537-019-0217-0>
- Davis, A. P., Grondin, C. J., Johnson, R. J., Sciaky, D., Wieggers, J., Wieggers, T. C., & Mattingly, C. J. (2021). Comparative Toxicogenomics Database (CTD): update 2021. *Nucleic Acids Research*, 49(D1), D1138–D1143. <https://doi.org/10.1093/nar/gkaa891>
- Davis, A. P., Wieggers, T. C., Johnson, R. J., Sciaky, D., Wieggers, J., & Mattingly, C. J. (2023). Comparative Toxicogenomics Database (CTD): update 2023. *Nucleic Acids Research*, 51(D1), D1257–D1262. <https://doi.org/10.1093/nar/gkac833>
- Di Lieto, E., Serra, A., Inkala, S. I., Saarimäki, L. A., Del Giudice, G., Fratello, M., Hautanen, V., Annala, M., Federico, A., & Greco, D. (2023). ESPERANTO: a GLP-fied sEmi-SuPERvised toxicogenomics meta-dAta curatioN TOol. *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btad405>
- Domenico, A., Nicola, G., Daniela, T., Fulvio, C., Nicola, A., & Orazio, N. (2020). De Novo Drug Design of Targeted Chemical Libraries Based on Artificial Intelligence

- and Pair-Based Multiobjective Optimization. *Journal of Chemical Information and Modeling*, 60(10), 4582–4593. <https://doi.org/10.1021/acs.jcim.0c00517>
- Dumbrava, E. I., & Meric-Bernstam, F. (2018). Personalized cancer therapy-leveraging a knowledge base for clinical decision-making. *Molecular Case Studies*, 4(2). <https://doi.org/10.1101/mcs.a001578>
- Du, Y., Cai, M., Xing, X., Ji, J., Yang, E., & Wu, J. (2021). PINA 3.0: mining cancer interactome. *Nucleic Acids Research*, 49(D1), D1351–D1357. <https://doi.org/10.1093/nar/gkaa1075>
- Ehrlinger, L., & WöB, W. (2016). Towards a Definition of Knowledge Graphs. In M. Martin, M. Cuquet, & E. Folmer (Eds.), *Proceedings of the Posters and Demos Track of the 12th International Conference on Semantic Systems - SEMANTiCS2016 and the 1st International Workshop on Semantic Change & Evolving Semantics (SuCCESS'16)* (Vol. 1695). CEUR-WS.
- Eidsaa, M., Stubbs, L., & Almaas, E. (2017). Comparative analysis of weighted gene co-expression networks in human and mouse. *Plos One*, 12(11), e0187611. <https://doi.org/10.1371/journal.pone.0187611>
- Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J. J., & Gardner, T. S. (2007). Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biology*, 5(1), e8. <https://doi.org/10.1371/journal.pbio.0050008>
- Fang, Y., Pan, X., & Shen, H.-B. (2023). De novo drug design by iterative multiobjective deep reinforcement learning with graph-based molecular quality assessment. *Bioinformatics*, 39(4). <https://doi.org/10.1093/bioinformatics/btad157>
- Fazekas, D., Koltai, M., Türei, D., Módos, D., Pálffy, M., Dúl, Z., Zsákai, L., Szalay-Bekő, M., Lenti, K., Farkas, I. J., Vellai, T., Csermely, P., & Korcsmáros, T. (2013).

- Signalink 2 - a signaling pathway resource with multi-layered regulatory networks. *BMC Systems Biology*, 7, 7. <https://doi.org/10.1186/1752-0509-7-7>
- Federico, A., Fratello, M., Scala, G., Möbus, L., Pavel, A., Del Giudice, G., Ceccarelli, M., Costa, V., Ciccodicola, A., Fortino, V., Serra, A., & Greco, D. (2022). Integrated Network Pharmacology Approach for Drug Combination Discovery: A Multi-Cancer Case Study. *Cancers*, 14(8). <https://doi.org/10.3390/cancers14082043>
- Federico, A., Pavel, A., Möbus, L., McKean, D., Del Giudice, G., Fortino, V., Niehues, H., Rastrick, J., Eyerich, K., Eyerich, S., van den Bogaard, E., Smith, C., Weidinger, S., de Rinaldis, E., & Greco, D. (2022). The integration of large-scale public data and network analysis uncovers molecular characteristics of psoriasis. *Human Genomics*, 16(1), 62. <https://doi.org/10.1186/s40246-022-00431-x>
- Federico, A., Serra, A., Ha, M. K., Kohonen, P., Choi, J.-S., Liampa, I., Nymark, P., Sanabria, N., Cattelani, L., Fratello, M., Kinaret, P. A. S., Jagiello, K., Puzyn, T., Melagraki, G., Gulumian, M., Afantitis, A., Sarimveis, H., Yoon, T.-H., Grafström, R., & Greco, D. (2020). Transcriptomics in toxicogenomics, part II: preprocessing and differential expression analysis for high quality data. *Nanomaterials (Basel, Switzerland)*, 10(5). <https://doi.org/10.3390/nano10050903>
- Ferri, P., Sáez, C., Félix-De Castro, A., Juan-Albarracín, J., Blanes-Selva, V., Sánchez-Cuesta, P., & García-Gómez, J. M. (2021). Deep ensemble multitask classification of emergency medical call incidents combining multimodal data improves emergency medical dispatch. *Artificial Intelligence in Medicine*, 117, 102088. <https://doi.org/10.1016/j.artmed.2021.102088>
- Fortino, V., Kinaret, P. A. S., Fratello, M., Serra, A., Saarimäki, L. A., Gallud, A., Gupta, G., Vales, G., Correia, M., Rasool, O., Ytterberg, J., Monopoli, M., Skoog, T., Ritchie, P., Moya, S., Vázquez-Campos, S., Handy, R., Grafström, R., Tran, L.,

... Greco, D. (2022). Biomarkers of nanomaterials hazard from multi-layer data. *Nature Communications*, *13*(1), 3798. <https://doi.org/10.1038/s41467-022-31609-5>

Franklin, J. M., Patorno, E., Desai, R. J., Glynn, R. J., Martin, D., Quinto, K., Pawar, A., Bessette, L. G., Lee, H., Garry, E. M., Gautam, N., & Schneeweiss, S. (2021). Emulating Randomized Clinical Trials With Nonrandomized Real-World Evidence Studies: First Results From the RCT DUPLICATE Initiative. *Circulation*, *143*(10), 1002–1013. <https://doi.org/10.1161/CIRCULATIONAHA.120.051718>

Fröhlich, H., Balling, R., Beerenwinkel, N., Kohlbacher, O., Kumar, S., Lengauer, T., Maathuis, M. H., Moreau, Y., Murphy, S. A., Przytycka, T. M., Rebhan, M., Röst, H., Schuppert, A., Schwab, M., Spang, R., Stekhoven, D., Sun, J., Weber, A., Ziemek, D., & Zupan, B. (2018). From hype to reality: data science enabling personalized medicine. *BMC Medicine*, *16*(1), 150. <https://doi.org/10.1186/s12916-018-1122-7>

Frolkis, A., Knox, C., Lim, E., Jewison, T., Law, V., Hau, D. D., Liu, P., Gautam, B., Ly, S., Guo, A. C., Xia, J., Liang, Y., Shrivastava, S., & Wishart, D. S. (2010). SMPDB: the small molecule pathway database. *Nucleic Acids Research*, *38*(Database issue), D480-7. <https://doi.org/10.1093/nar/gkp1002>

Funtanilla, V. D., Candidate, P., Caliendo, T., & Hilas, O. (2019). Continuous glucose monitoring: A review of available systems. *P & T: A Peer-Reviewed Journal for Formulary Management*, *44*(9), 550–553.

Gagliano Taliun, S. A., VandeHaar, P., Boughton, A. P., Welch, R. P., Taliun, D., Schmidt, E. M., Zhou, W., Nielsen, J. B., Willer, C. J., Lee, S., Fritsche, L. G., Boehnke, M., & Abecasis, G. R. (2020). Exploring and visualizing large-scale genetic associations by using PheWeb. *Nature Genetics*, *52*(6), 550–552. <https://doi.org/10.1038/s41588-020-0622-5>

- Galeano, D., Li, S., Gerstein, M., & Paccanaro, A. (2020). Predicting the frequencies of drug side effects. *Nature Communications*, *11*(1), 4575.  
<https://doi.org/10.1038/s41467-020-18305-y>
- Galeano, D., & Paccanaro, A. (2022). Machine learning prediction of side effects for drugs in clinical trials. *Cell Reports Methods*, *2*(12), 100358.  
<https://doi.org/10.1016/j.crmeth.2022.100358>
- Gandhi, M., Aweeka, F., Greenblatt, R. M., & Blaschke, T. F. (2004). Sex differences in pharmacokinetics and pharmacodynamics. *Annual Review of Pharmacology and Toxicology*, *44*, 499–523.  
<https://doi.org/10.1146/annurev.pharmtox.44.101802.121453>
- Ganesan, S., Anaimalai Thirumurthi, N., Raghunath, A., Vijayakumar, S., & Perumal, E. (2016). Acute and sub-lethal exposure to copper oxide nanoparticles causes oxidative stress and teratogenicity in zebrafish embryos. *Journal of Applied Toxicology*, *36*(4), 554–567. <https://doi.org/10.1002/jat.3224>
- Gan, J.-H., Liu, J.-X., Liu, Y., Chen, S.-W., Dai, W.-T., Xiao, Z.-X., & Cao, Y. (2023). DrugRep: an automatic virtual screening server for drug repurposing. *Acta Pharmacologica Sinica*, *44*(4), 888–896. <https://doi.org/10.1038/s41401-022-00996-2>
- Gan, Y., Hu, X., Zou, G., Yan, C., & Xu, G. (2022). Inferring Gene Regulatory Networks From Single-Cell Transcriptomic Data Using Bidirectional RNN. *Frontiers in Oncology*, *12*, 899825. <https://doi.org/10.3389/fonc.2022.899825>
- Gao, S., Han, L., Luo, D., Liu, G., Xiao, Z., Shan, G., Zhang, Y., & Zhou, W. (2021). Modeling drug mechanism of action with large scale gene-expression profiles using GPAR, an artificial intelligence platform. *BMC Bioinformatics*, *22*(1), 17.  
<https://doi.org/10.1186/s12859-020-03915-6>

- Gao, Z., Ding, P., & Xu, R. (2022). KG-Predict: A knowledge graph computational framework for drug repurposing. *Journal of Biomedical Informatics*, *132*, 104133. <https://doi.org/10.1016/j.jbi.2022.104133>
- Gardiner, L.-J., Carrieri, A. P., Wilshaw, J., Checkley, S., Pyzer-Knapp, E. O., & Krishna, R. (2020). Using human in vitro transcriptome analysis to build trustworthy machine learning models for prediction of animal drug toxicity. *Scientific Reports*, *10*(1), 9522. <https://doi.org/10.1038/s41598-020-66481-0>
- Ge, Y., Tian, T., Huang, S., Wan, F., Li, J., Li, S., Wang, X., Yang, H., Hong, L., Wu, N., Yuan, E., Luo, Y., Cheng, L., Hu, C., Lei, Y., Shu, H., Feng, X., Jiang, Z., Wu, Y., ... Zeng, J. (2021). An integrative drug repositioning framework discovered a potential therapeutic agent targeting COVID-19. *Signal Transduction and Targeted Therapy*, *6*(1), 165. <https://doi.org/10.1038/s41392-021-00568-6>
- Giarratano, Y., Pavel, A., Lian, J., Andreeva, R., Fontanella, A., Sarkar, R., Reid, L. J., Forbes, S., Pugh, D., Farrah, T. E., Dhaun, N., Dhillon, B., MacGillivray, T., & Bernabeu, M. O. (2020). A framework for the discovery of retinal biomarkers in optical coherence tomography angiography (OCTA). In H. Fu, M. K. Garvin, T. MacGillivray, Y. Xu, & Y. Zheng (Eds.), *Ophthalmic Medical Image Analysis: 7th International Workshop, OMLA 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 8, 2020, Proceedings* (Vol. 12069, pp. 155–164). Springer International Publishing. [https://doi.org/10.1007/978-3-030-63419-3\\_16](https://doi.org/10.1007/978-3-030-63419-3_16)
- Glez-Peña, D., Alvarez, R., Díaz, F., & Fdez-Riverola, F. (2009). DFP: a Bioconductor package for fuzzy profile identification and gene reduction of microarray data. *BMC Bioinformatics*, *10*, 37. <https://doi.org/10.1186/1471-2105-10-37>



- Goetz, L. H., & Schork, N. J. (2018). Personalized medicine: motivation, challenges, and progress. *Fertility and Sterility*, *109*(6), 952–963.  
<https://doi.org/10.1016/j.fertnstert.2018.05.006>
- Goldstein, J. A., Bastarache, L. A., Denny, J. C., Roden, D. M., Pulley, J. M., & Aronoff, D. M. (2018). Calcium channel blockers as drug repurposing candidates for gestational diabetes: Mining large scale genomic and electronic health records data to repurpose medications. *Pharmacological Research*, *130*, 44–51.  
<https://doi.org/10.1016/j.phrs.2018.02.013>
- Gordon, D. E., Jang, G. M., Bouhaddou, M., Xu, J., Obernier, K., White, K. M., O’Meara, M. J., Rezelj, V. V., Guo, J. Z., Swaney, D. L., Tummino, T. A., Hüttenhain, R., Kaake, R. M., Richards, A. L., Tutuncuoglu, B., Foussard, H., Batra, J., Haas, K., Modak, M., ... Krogan, N. J. (2020). A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature*, *583*(7816), 459–468.  
<https://doi.org/10.1038/s41586-020-2286-9>
- Gosak, M., Markovič, R., Dolensek, J., Slak Rupnik, M., Marhl, M., Stožer, A., & Perc, M. (2018). Network science of biological systems at different scales: A review. *Physics of Life Reviews*, *24*, 118–135. <https://doi.org/10.1016/j.plrev.2017.11.003>
- Gosztolai, A., & Ramdya, P. (2022). Connecting the dots in ethology: applying network theory to understand neural and animal collectives. *Current Opinion in Neurobiology*, *73*, 102532. <https://doi.org/10.1016/j.conb.2022.102532>
- Grover, A., & Leskovec, J. (2016). node2vec: Scalable Feature Learning for Networks. *KDD: Proceedings / International Conference on Knowledge Discovery & Data Mining. International Conference on Knowledge Discovery & Data Mining, 2016*, 855–864.  
<https://doi.org/10.1145/2939672.2939754>

- GTEX Consortium. (2013). The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, 45(6), 580–585. <https://doi.org/10.1038/ng.2653>
- Gu, Y., Zalkikar, A., Liu, M., Kelly, L., Hall, A., Daly, K., & Ward, T. (2021). Predicting medication adherence using ensemble learning and deep learning models with large scale healthcare data. *Scientific Reports*, 11(1), 18961. <https://doi.org/10.1038/s41598-021-98387-w>
- Günther, W. A., Rezazade Mehrizi, M. H., Huysman, M., & Feldberg, F. (2017). Debating big data: A literature review on realizing value from big data. *The Journal of Strategic Information Systems*. <https://doi.org/10.1016/j.jsis.2017.07.003>
- Guo, J., Ning, Y., Su, Z., Guo, L., & Gu, Y. (2022). Identification of hub genes and regulatory networks in histologically unstable carotid atherosclerotic plaque by bioinformatics analysis. *BMC Medical Genomics*, 15(1), 145. <https://doi.org/10.1186/s12920-022-01257-1>
- Gupta, S., Modgil, S., & Gunasekaran, A. (2019). Big data in lean six sigma: a review and further research directions. *International Journal of Production Research*, 1–23. <https://doi.org/10.1080/00207543.2019.1598599>
- Gutiérrez-Sacristán, A., Bravo, À., Portero-Tresserra, M., Valverde, O., Armario, A., Blanco-Gandía, M. C., Farré, A., Fernández-Ibarrondo, L., Fonseca, F., Giraldo, J., Leis, A., Mané, A., Mayer, M. A., Montagud-Romero, S., Nadal, R., Ortiz, J., Pavon, F. J., Perez, E. J., Rodríguez-Arias, M., ... Furlong, L. I. (2017). Text mining and expert curation to develop a database on psychiatric diseases and their genes. *Database: The Journal of Biological Databases and Curation*, 2017. <https://doi.org/10.1093/database/bax043>
- Gutiérrez-Sacristán, A., Grosdidier, S., Valverde, O., Torrens, M., Bravo, À., Piñero, J., Sanz, F., & Furlong, L. I. (2015). PsyGeNET: a knowledge platform on psychiatric

- disorders and their genes. *Bioinformatics*, 31(18), 3075–3077.  
<https://doi.org/10.1093/bioinformatics/btv301>
- Hagberg, A. A., Schult, D. A., & Swart, P. J. (2008, August). Exploring network structure, dynamics, and function using networkx. *Proceedings of the 7th Python in Science Conference (SciPy2008)*. SciPy2008.
- Haghanifar, A., Majdabadi, M. M., & Ko, S.-B. (2020). Automated Teeth Extraction from Dental Panoramic X-Ray Images using Genetic Algorithm. *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*, 1–5.  
<https://doi.org/10.1109/ISCAS45731.2020.9180937>
- Hamming, R. W. (1980). *Coding and Information Theory*. Prentice-Hall Inc.
- Han, H., Cho, J.-W., Lee, S., Yun, A., Kim, H., Bae, D., Yang, S., Kim, C. Y., Lee, M., Kim, E., Lee, S., Kang, B., Jeong, D., Kim, Y., Jeon, H.-N., Jung, H., Nam, S., Chung, M., Kim, J.-H., & Lee, I. (2018). TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Research*, 46(D1), D380–D386. <https://doi.org/10.1093/nar/gkx1013>
- Harding, S. D., Armstrong, J. F., Faccenda, E., Southan, C., Alexander, S. P. H., Davenport, A. P., Pawson, A. J., Spedding, M., Davies, J. A., & NC-IUPHAR. (2022). The IUPHAR/BPS guide to PHARMACOLOGY in 2022: curating pharmacology for COVID-19, malaria and antibacterials. *Nucleic Acids Research*, 50(D1), D1282–D1294. <https://doi.org/10.1093/nar/gkab1010>
- Hasankhani, A., Bahrami, A., Sheybani, N., Aria, B., Hemati, B., Fatehi, F., Ghaem Maghami Farahani, H., Javanmard, G., Rezaee, M., Kastelic, J. P., & Barkema, H. W. (2021). Differential Co-Expression Network Analysis Reveals Key Hub-High Traffic Genes as Potential Therapeutic Targets for COVID-19 Pandemic. *Frontiers in Immunology*, 12, 789317. <https://doi.org/10.3389/fimmu.2021.789317>

- Hayes, W., Sun, K., & Pržulj, N. (2013). Graphlet-based measures are suitable for biological network comparison. *Bioinformatics*, *29*(4), 483–491.  
<https://doi.org/10.1093/bioinformatics/bts729>
- Hernán, M. A., & Robins, J. M. (2016). Using big data to emulate a target trial when a randomized trial is not available. *American Journal of Epidemiology*, *183*(8), 758–764.  
<https://doi.org/10.1093/aje/kwv254>
- Hogan, A., Blomqvist, E., Cochez, M., D'amato, C., Melo, G. D., Gutierrez, C., Kirrane, S., Gayo, J. E. L., Navigli, R., Neumaier, S., Ngomo, A.-C. N., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., & Zimmermann, A. (2021). Knowledge Graphs. *ACM Computing Surveys*, *54*(4), 1–37.  
<https://doi.org/10.1145/3447772>
- Hornos Carneiro, M. F., & Barbosa, F. (2016). Gold nanoparticles: A critical review of therapeutic applications and toxicological aspects. *Journal of Toxicology and Environmental Health. Part B, Critical Reviews*, *19*(3–4), 129–148.  
<https://doi.org/10.1080/10937404.2016.1168762>
- Hou, J., & Gao, T. (2021). Explainable DCNN based chest X-ray image analysis and classification for COVID-19 pneumonia detection. *Scientific Reports*, *11*(1), 16071.  
<https://doi.org/10.1038/s41598-021-95680-6>
- Huang, H.-Y., Lin, Y.-C.-D., Cui, S., Huang, Y., Tang, Y., Xu, J., Bao, J., Li, Y., Wen, J., Zuo, H., Wang, W., Li, J., Ni, J., Ruan, Y., Li, L., Chen, Y., Xie, Y., Zhu, Z., Cai, X., ... Huang, H.-D. (2022). miRTarBase update 2022: an informative resource for experimentally validated miRNA-target interactions. *Nucleic Acids Research*, *50*(D1), D222–D230. <https://doi.org/10.1093/nar/gkab1079>
- Hughes, L. D., Tsueng, G., DiGiovanna, J., Horvath, T. D., Rasmussen, L. V., Savidge, T. C., Stoeger, T., Turkarslan, S., Wu, Q., Wu, C., Su, A. I., Pache, L., & NIAID

- Systems Biology Data Dissemination Working Group. (2023). Addressing barriers in FAIR data practices for biomedical data. *Scientific Data*, *10*(1), 98.  
<https://doi.org/10.1038/s41597-023-01969-8>
- Hu, Jiajing, Lepore, R., Dobson, R. J. B., Al-Chalabi, A., M Bean, D., & Iacoangeli, A. (2021). DGLinker: flexible knowledge-graph prediction of disease-gene associations. *Nucleic Acids Research*, *49*(W1), W153–W161. <https://doi.org/10.1093/nar/gkab449>
- Hu, Jianfei, Rho, H.-S., Newman, R. H., Zhang, J., Zhu, H., & Qian, J. (2014). PhosphoNetworks: a database for human phosphorylation networks. *Bioinformatics*, *30*(1), 141–142. <https://doi.org/10.1093/bioinformatics/btt627>
- Igarashi, Y., Nakatsu, N., Yamashita, T., Ono, A., Ohno, Y., Urushidani, T., & Yamada, H. (2015). Open TG-GATEs: a large-scale toxicogenomics database. *Nucleic Acids Research*, *43*(Database issue), D921-7.  
<https://doi.org/10.1093/nar/gku955>
- Irwin, J. J., Tang, K. G., Young, J., Dandarchuluun, C., Wong, B. R., Khurelbaatar, M., Moroz, Y. S., Mayfield, J., & Sayle, R. A. (2020). ZINC20-A Free Ultralarge-Scale Chemical Database for Ligand Discovery. *Journal of Chemical Information and Modeling*, *60*(12), 6065–6073. <https://doi.org/10.1021/acs.jcim.0c00675>
- Ivanov, M., Barragan, I., & Ingelman-Sundberg, M. (2014). Epigenetic mechanisms of importance for drug treatment. *Trends in Pharmacological Sciences*, *35*(8), 384–396.  
<https://doi.org/10.1016/j.tips.2014.05.004>
- Jaber, M. I., Song, B., Taylor, C., Vaske, C. J., Benz, S. C., Rabizadeh, S., Soon-Shiong, P., & Szeto, C. W. (2020). A deep learning image-based intrinsic molecular subtype classifier of breast tumors reveals tumor heterogeneity that may affect survival. *Breast Cancer Research*, *22*(1), 12. <https://doi.org/10.1186/s13058-020-1248-3>

- Jaccard, P. (1908). Nouvelles Recherches Sur La Distribution Florale. . *Ulletin de La Société Vaudoise Des Sciences Naturelles*, *44*, 223–270.
- Jafari, A., Babajani, A., Sarrami Forooshani, R., Yazdani, M., & Rezaei-Tavirani, M. (2022). Clinical applications and anticancer effects of antimicrobial peptides: from bench to bedside. *Frontiers in Oncology*, *12*, 819563.  
<https://doi.org/10.3389/fonc.2022.819563>
- Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., Sidiropoulos, K., Cook, J., Gillespie, M., Haw, R., Loney, F., May, B., Milacic, M., Rothfels, K., Sevilla, C., Shamovsky, V., Shorsler, S., Varusai, T., Weiser, J., ... D'Eustachio, P. (2020). The Reactome Pathway Knowledgebase. *Nucleic Acids Research*, *48*(D1), D498–D503. <https://doi.org/10.1093/nar/gkz1031>
- Jensen, P. B., Jensen, L. J., & Brunak, S. (2012). Mining electronic health records: Towards better research applications and clinical care. *Nature Reviews. Genetics*, *13*(6), 395–405. <https://doi.org/10.1038/nrg3208>
- Jeong, H., Qian, X., & Yoon, B.-J. (2016). Effective comparative analysis of protein-protein interaction networks by measuring the steady-state network flow using a Markov model. *BMC Bioinformatics*, *17*(Suppl 13), 395.  
<https://doi.org/10.1186/s12859-016-1215-2>
- Jewison, T., Su, Y., Disfany, F. M., Liang, Y., Knox, C., Maciejewski, A., Poelzer, J., Huynh, J., Zhou, Y., Arndt, D., Djoumbou, Y., Liu, Y., Deng, L., Guo, A. C., Han, B., Pon, A., Wilson, M., Rafatnia, S., Liu, P., & Wishart, D. S. (2014). SMPDB 2.0: big improvements to the Small Molecule Pathway Database. *Nucleic Acids Research*, *42*(Database issue), D478-84. <https://doi.org/10.1093/nar/gkt1067>
- Ji, S., Pan, S., Cambria, E., Marttinen, P., & Yu, P. S. (2022). A survey on knowledge graphs: representation, acquisition, and applications. *IEEE Transactions on Neural*

- Networks and Learning Systems*, 33(2), 494–514.  
<https://doi.org/10.1109/TNNLS.2021.3070843>
- Ji, Z., Zhang, D., Li, L., Shen, X., Deng, X., Dong, L., Wu, M., & Liu, Y. (2009). The hepatotoxicity of multi-walled carbon nanotubes in mice. *Nanotechnology*, 20(44), 445101. <https://doi.org/10.1088/0957-4484/20/44/445101>
- Kamath, A., Mahalingam, A., & Brauker, J. (2010). Methods of evaluating the utility of continuous glucose monitor alerts. *Journal of Diabetes Science and Technology*, 4(1), 57–66. <https://doi.org/10.1177/193229681000400108>
- Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1), 27–30. <https://doi.org/10.1093/nar/28.1.27>
- Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., & Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45(D1), D353–D361. <https://doi.org/10.1093/nar/gkw1092>
- Karim, Md. R., Cochez, M., Jares, J. B., Uddin, M., Beyan, O., & Decker, S. (2019). Drug-Drug Interaction Prediction Based on Knowledge Graph Embeddings and Convolutional-LSTM Network. *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics - BCB '19*, 113–123.  
<https://doi.org/10.1145/3307339.3342161>
- Kelleher, K. J., Sheils, T. K., Mathias, S. L., Yang, J. J., Metzger, V. T., Siramshetty, V. B., Nguyen, D.-T., Jensen, L. J., Vidović, D., Schürer, S. C., Holmes, J., Sharma, K. R., Pillai, A., Bologna, C. G., Edwards, J. S., Mathé, E. A., & Oprea, T. I. (2023). Pharos 2023: an integrated resource for the understudied human proteome. *Nucleic Acids Research*, 51(D1), D1405–D1416. <https://doi.org/10.1093/nar/gkac1033>
- Kepp, O., Menger, L., Vacchelli, E., Adjemian, S., Martins, I., Ma, Y., Sukkurwala, A. Q., Michaud, M., Galluzzi, L., Zitvogel, L., & Kroemer, G. (2012). Anticancer

- activity of cardiac glycosides: At the frontier between cell-autonomous and immunological effects. *Oncoimmunology*, 1(9), 1640–1642.  
<https://doi.org/10.4161/onci.21684>
- Khare, R., Li, J., & Lu, Z. (2014). LabeledIn: cataloging labeled indications for human drugs. *Journal of Biomedical Informatics*, 52, 448–456.  
<https://doi.org/10.1016/j.jbi.2014.08.004>
- Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., Zaslavsky, L., Zhang, J., & Bolton, E. E. (2023). PubChem 2023 update. *Nucleic Acids Research*, 51(D1), D1373–D1380.  
<https://doi.org/10.1093/nar/gkac956>
- Kinaret, P., Marwah, V., Fortino, V., Ilves, M., Wolff, H., Ruokolainen, L., Auvinen, P., Savolainen, K., Alenius, H., & Greco, D. (2017). Network analysis reveals similar transcriptomic responses to intrinsic properties of carbon nanomaterials in vitro and in vivo. *ACS Nano*, 11(4), 3786–3796. <https://doi.org/10.1021/acsnano.6b08650>
- Kinaret, P., Del Giudice, G., & Greco, D. (2020). Covid-19 acute responses and possible long term consequences: What nanotoxicology can teach us. *Nano Today*, 35, 100945. <https://doi.org/10.1016/j.nantod.2020.100945>
- Kinaret, P., Serra, A., Federico, A., Kohonen, P., Nymark, P., Liampa, I., Ha, M. K., Choi, J.-S., Jagiello, K., Sanabria, N., Melagraki, G., Cattelani, L., Fratello, M., Sarimveis, H., Afantitis, A., Yoon, T.-H., Gulumian, M., Grafström, R., Puzyn, T., & Greco, D. (2020). Transcriptomics in toxicogenomics, part I: experimental design, technologies, publicly available data, and regulatory aspects. *Nanomaterials (Basel, Switzerland)*, 10(4). <https://doi.org/10.3390/nano10040750>
- Kingston, D. G. I. (2009). Tubulin-interactive natural products as anticancer agents. *Journal of Natural Products*, 72(3), 507–515. <https://doi.org/10.1021/np800568j>



- Knapen, D., Angrish, M. M., Fortin, M. C., Katsiadaki, I., Leonard, M., Margiotta-Casaluci, L., Munn, S., O'Brien, J. M., Pollesch, N., Smith, L. C., Zhang, X., & Villeneuve, D. L. (2018). Adverse outcome pathway networks I: Development and applications. *Environmental Toxicology and Chemistry*, *37*(6), 1723–1733.  
<https://doi.org/10.1002/etc.4125>
- Koenig, N., Almunia, C., Bonnal-Conduzorgues, A., Armengaud, J., Chaumot, A., Geffard, O., & Esposti, D. D. (2021). Co-expression network analysis identifies novel molecular pathways associated with cadmium and pyriproxyfen testicular toxicity in *Gammarus fossarum*. *Aquatic Toxicology*, *235*, 105816.  
<https://doi.org/10.1016/j.aquatox.2021.105816>
- Köhler, S., Gargano, M., Matentzoglou, N., Carmody, L. C., Lewis-Smith, D., Vasilevsky, N. A., Danis, D., Balagura, G., Baynam, G., Brower, A. M., Callahan, T. J., Chute, C. G., Est, J. L., Galer, P. D., Ganesan, S., Griese, M., Haimel, M., Pazmandi, J., Hanauer, M., ... Robinson, P. N. (2021). The human phenotype ontology in 2021. *Nucleic Acids Research*, *49*(D1), D1207–D1217.  
<https://doi.org/10.1093/nar/gkaa1043>
- Kohonen, P., Parkkinen, J. A., Willighagen, E. L., Ceder, R., Wennerberg, K., Kaski, S., & Grafström, R. C. (2017). A transcriptomics data-driven gene space accurately predicts liver cytopathology and drug-induced liver injury. *Nature Communications*, *8*, 15932. <https://doi.org/10.1038/ncomms15932>
- Kola, A., Lamponi, S., Currò, F., & Valensin, D. (2023). A Comparative Study between Lycorine and Galantamine Abilities to Interact with AMYLOID  $\beta$  and Reduce In Vitro Neurotoxicity. *International Journal of Molecular Sciences*, *24*(3).  
<https://doi.org/10.3390/ijms24032500>

- Koutrouli, M., Karatzas, E., Paez-Espino, D., & Pavlopoulos, G. A. (2020). A guide to conquer the biological network era using graph theory. *Frontiers in Bioengineering and Biotechnology*, *8*, 34. <https://doi.org/10.3389/fbioe.2020.00034>
- Kovačević, Ž., Gurbeta Pokvić, L., Spahić, L., & Badnjević, A. (2020). Prediction of medical device performance using machine learning techniques: infant incubator case study. *Health and Technology*, *10*(1), 151–155. <https://doi.org/10.1007/s12553-019-00386-5>
- Kruger, D. F., Edelman, S. V., Hinnen, D. A., & Parkin, C. G. (2019). Reference guide for integrating continuous glucose monitoring into clinical practice. *The Diabetes Educator*, *45*(1\_suppl), 3S-20S. <https://doi.org/10.1177/0145721718818066>
- Kuhn, M., Letunic, I., Jensen, L. J., & Bork, P. (2016). The SIDER database of drugs and side effects. *Nucleic Acids Research*, *44*(D1), D1075-9. <https://doi.org/10.1093/nar/gkv1075>
- Kuntal, B. K., Dutta, A., & Mande, S. S. (2016). CompNet: a GUI based tool for comparison of multiple biological interaction networks. *BMC Bioinformatics*, *17*(1), 185. <https://doi.org/10.1186/s12859-016-1013-x>
- Kurnit, K. C., Bailey, A. M., Zeng, J., Johnson, A. M., Shufean, M. A., Brusco, L., Litzemberger, B. C., Sánchez, N. S., Khotskaya, Y. B., Holla, V., Simpson, A., Mills, G. B., Mendelsohn, J., Bernstam, E., Shaw, K., & Meric-Bernstam, F. (2017). “personalized cancer therapy”: A publicly available precision oncology resource. *Cancer Research*, *77*(21), e123–e126. <https://doi.org/10.1158/0008-5472.CAN-17-0341>
- Kurt, I., Ture, M., & Kurum, A. T. (2008). Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting

- coronary artery disease. *Expert Systems with Applications*, 34(1), 366–374.  
<https://doi.org/10.1016/j.eswa.2006.09.004>
- Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., Lerner, J., Brunet, J.-P., Subramanian, A., Ross, K. N., Reich, M., Hieronymus, H., Wei, G., Armstrong, S. A., Haggarty, S. J., Clemons, P. A., Wei, R., Carr, S. A., Lander, E. S., & Golub, T. R. (2006). The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313(5795), 1929–1935.  
<https://doi.org/10.1126/science.1132939>
- Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., Karapetyan, K., Katz, K., Liu, C., Maddipatla, Z., Malheiro, A., McDaniel, K., Ovetsky, M., Riley, G., Zhou, G., ... Maglott, D. R. (2018). ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, 46(D1), D1062–D1067.  
<https://doi.org/10.1093/nar/gkx1153>
- Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9, 559. <https://doi.org/10.1186/1471-2105-9-559>
- Lao, N., Mitchell, T., & Cohen, W. (2011). Random walk inference and learning in a large scale knowledge base. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 529–539.
- Leaman, R., Khare, R., & Lu, Z. (2015). Challenges in clinical natural language processing for automated disorder normalization. *Journal of Biomedical Informatics*, 57, 28–37. <https://doi.org/10.1016/j.jbi.2015.07.010>
- Lee, S.-I., Celik, S., Logsdon, B. A., Lundberg, S. M., Martins, T. J., Oehler, V. G., Estey, E. H., Miller, C. P., Chien, S., Dai, J., Saxena, A., Blau, C. A., & Becker, P. S.

- (2018). A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia. *Nature Communications*, 9(1), 42.  
<https://doi.org/10.1038/s41467-017-02465-5>
- Leonelli, S. (2014). What difference does quantity make? on the epistemology of big data in biology. *Big Data & Society*, 1(1).  
<https://doi.org/10.1177/2053951714534395>
- Leonelli, S. (2019). The challenges of big data biology. *ELife*, 8.  
<https://doi.org/10.7554/eLife.47381>
- Leskovec, J., Backstrom, L., Kumar, R., & Tomkins, A. (2008). Microscopic evolution of social networks. *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 08*, 462.  
<https://doi.org/10.1145/1401890.1401948>
- Liang, X., Fu, Y., Qu, L., Zhang, P., & Chen, Y. (2023). Prediction of drug side effects with transductive matrix co-completion. *Bioinformatics*, 39(1).  
<https://doi.org/10.1093/bioinformatics/btad006>
- Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., & Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Systems*, 1(6), 417–425. <https://doi.org/10.1016/j.cels.2015.12.004>
- Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., & Mesirov, J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, 27(12), 1739–1740. <https://doi.org/10.1093/bioinformatics/btr260>
- Linhares, C. D. G., Ponciano, J. R., Pereira, F. S. F., Rocha, L. E. C., Paiva, J. G. S., & Travençolo, B. A. N. (2020). Visual analysis for evaluation of community detection algorithms. *Multimedia Tools and Applications*, 79(25–26), 17645–17667.  
<https://doi.org/10.1007/s11042-020-08700-4>

- Liu, A., Seal, S., Yang, H., & Bender, A. (2023). Using chemical and biological data to predict drug toxicity. *SLAS Discovery*, 28(3), 53–64.  
<https://doi.org/10.1016/j.slasd.2022.12.003>
- Liu, H., Zhang, W., Zou, B., Wang, J., Deng, Y., & Deng, L. (2020). DrugCombDB: a comprehensive database of drug combinations toward the discovery of combinatorial therapy. *Nucleic Acids Research*, 48(D1), D871–D881.  
<https://doi.org/10.1093/nar/gkz1007>
- Liu, M., Zhang, D., & Shen, D. (2015). Inherent Structure-Guided Multi-view Learning for Alzheimer’s Disease and Mild Cognitive Impairment Classification. *Machine Learning in Medical Imaging. MLMI (Workshop), Author*, 9352, 296–303.  
[https://doi.org/10.1007/978-3-319-24888-2\\_36](https://doi.org/10.1007/978-3-319-24888-2_36)
- Liu, W., Tu, W., Li, L., Liu, Y., Wang, S., Li, L., Tao, H., & He, H. (2018). Revisiting Connectivity Map from a gene co-expression network analysis. *Experimental and Therapeutic Medicine*, 16(2), 493–500. <https://doi.org/10.3892/etm.2018.6275>
- Liu, X., Wang, S., Meng, F., Wang, J., Zhang, Y., Dai, E., Yu, X., Li, X., & Jiang, W. (2013). SM2miR: a database of the experimentally validated small molecules’ effects on microRNA expression. *Bioinformatics*, 29(3), 409–411.  
<https://doi.org/10.1093/bioinformatics/bts698>
- Liu, Y., Gu, H.-Y., Zhu, J., Niu, Y.-M., Zhang, C., & Guo, G.-L. (2019). Identification of Hub Genes and Key Pathways Associated With Bipolar Disorder Based on Weighted Gene Co-expression Network Analysis. *Frontiers in Physiology*, 10, 1081.  
<https://doi.org/10.3389/fphys.2019.01081>
- Liu, Z., Huang, R., Roberts, R., & Tong, W. (2019). Toxicogenomics: A 2020 vision. *Trends in Pharmacological Sciences*, 40(2), 92–103.  
<https://doi.org/10.1016/j.tips.2018.12.001>

- Li, A., & Bergan, R. C. (2020). Clinical trial design: Past, present, and future in the context of big data and precision medicine. *Cancer*, *126*(22), 4838–4846.  
<https://doi.org/10.1002/cncr.33205>
- Li, M., Zou, D., Li, Z., Gao, R., Sang, J., Zhang, Y., Li, R., Xia, L., Zhang, T., Niu, G., Bao, Y., & Zhang, Z. (2019). EWAS Atlas: a curated knowledgebase of epigenome-wide association studies. *Nucleic Acids Research*, *47*(D1), D983–D988.  
<https://doi.org/10.1093/nar/gky1027>
- Li, W., Wang, L., Wu, Y., Yuan, Z., & Zhou, J. (2020). Weighted gene co-expression network analysis to identify key modules and hub genes associated with atrial fibrillation. *International Journal of Molecular Medicine*, *45*(2), 401–416.  
<https://doi.org/10.3892/ijmm.2019.4416>
- Lobentanzer, S., Aloy, P., Baumbach, J., Bohar, B., Carey, V. J., Charoentong, P., Danhauser, K., Doğan, T., Dreo, J., Dunham, I., Farr, E., Fernandez-Torras, A., Gyori, B. M., Hartung, M., Hoyt, C. T., Klein, C., Korcsmaros, T., Maier, A., Mann, M., ... Saez-Rodriguez, J. (2023). Democratizing knowledge representation with BioCypher. *Nature Biotechnology*. <https://doi.org/10.1038/s41587-023-01848-y>
- Locke, S., Bashall, A., Al-Adely, S., Moore, J., Wilson, A., & Kitchen, G. B. (2021). Natural language processing in medicine: A review. *Trends in Anaesthesia and Critical Care*, *38*, 4–9. <https://doi.org/10.1016/j.tacc.2021.02.007>
- Lo, Y.-C., Rensi, S. E., Torng, W., & Altman, R. B. (2018). Machine learning in chemoinformatics and drug discovery. *Drug Discovery Today*, *23*(8), 1538–1546.  
<https://doi.org/10.1016/j.drudis.2018.05.010>
- López, Y., Nakai, K., & Patil, A. (2015). HitPredict version 4: comprehensive reliability scoring of physical protein-protein interactions from more than 100 species.

*Database: The Journal of Biological Databases and Curation, 2015.*

<https://doi.org/10.1093/database/bav117>

- Luck, K., Kim, D.-K., Lambourne, L., Spirohn, K., Begg, B. E., Bian, W., Brignall, R., Cafarelli, T., Campos-Laborie, F. J., Charloteaux, B., Choi, D., Coté, A. G., Daley, M., Deimling, S., Desbuleux, A., Dricot, A., Gebbia, M., Hardy, M. F., Kishore, N., ... Calderwood, M. A. (2020). A reference map of the human binary protein interactome. *Nature*, *580*(7803), 402–408. <https://doi.org/10.1038/s41586-020-2188-x>
- Luo, Y., Hitz, B. C., Gabdank, I., Hilton, J. A., Kagda, M. S., Lam, B., Myers, Z., Sud, P., Jou, J., Lin, K., Baymuradov, U. K., Graham, K., Litton, C., Miyasato, S. R., Strattan, J. S., Jolanki, O., Lee, J.-W., Tanaka, F. Y., Adenekan, P., ... Cherry, J. M. (2020). New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Research*, *48*(D1), D882–D889. <https://doi.org/10.1093/nar/gkz1062>
- Maglott, D., Ostell, J., Pruitt, K. D., & Tatusova, T. (2011). Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research*, *39*(Database issue), D52-7. <https://doi.org/10.1093/nar/gkq1237>
- Majolo, F., de Oliveira Becker Delwing, L. K., Marmitt, D. J., Bustamante-Filho, I. C., & Goettert, M. I. (2019). Medicinal plants and bioactive natural compounds for cancer treatment: Important advances for drug discovery. *Phytochemistry Letters*, *31*, 196–207. <https://doi.org/10.1016/j.phytol.2019.04.003>
- Makagon, M. M., McCowan, B., & Mench, J. A. (2012). How can social network analysis contribute to social behavior research in applied ethology? *Applied Animal Behaviour Science*, *138*(3–4). <https://doi.org/10.1016/j.applanim.2012.02.003>

- Maloney, M. P., Coley, C. W., Genheden, S., Carson, N., Helquist, P., Norrby, P.-O., & Wiest, O. (2023). Negative data in data sets for machine learning training. *The Journal of Organic Chemistry*, *88*(9), 5239–5241. <https://doi.org/10.1021/acs.joc.3c00844>
- Malone, J., Holloway, E., Adamusiak, T., Kapushesky, M., Zheng, J., Kolesnikov, N., Zhukova, A., Brazma, A., & Parkinson, H. (2010). Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics*, *26*(8), 1112–1118. <https://doi.org/10.1093/bioinformatics/btq099>
- Manczinger, M., Bodnár, V. Á., Papp, B. T., Bolla, S. B., Szabó, K., Balázs, B., Csányi, E., Szél, E., Erős, G., & Kemény, L. (2018). Drug Repurposing by Simulating Flow Through Protein-Protein Interaction Networks. *Clinical Pharmacology and Therapeutics*, *103*(3), 511–520. <https://doi.org/10.1002/cpt.769>
- Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., Allison, K. R., DREAM5 Consortium, Kellis, M., Collins, J. J., & Stolovitzky, G. (2012). Wisdom of crowds for robust gene network inference. *Nature Methods*, *9*(8), 796–804. <https://doi.org/10.1038/nmeth.2016>
- Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., & Califano, A. (2006). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, *7* Suppl 1(Suppl 1), S7. <https://doi.org/10.1186/1471-2105-7-S1-S7>
- Martens, M., Ammar, A., Riutta, A., Waagmeester, A., Slenter, D. N., Hanspers, K., A Miller, R., Digles, D., Lopes, E. N., Ehrhart, F., Dupuis, L. J., Winckers, L. A., Coort, S. L., Willighagen, E. L., Evelo, C. T., Pico, A. R., & Kutmon, M. (2021). WikiPathways: connecting communities. *Nucleic Acids Research*, *49*(D1), D613–D621. <https://doi.org/10.1093/nar/gkaa1024>



- Martha, V.-S., Liu, Z., Guo, L., Su, Z., Ye, Y., Fang, H., Ding, D., Tong, W., & Xu, X. (2011). Constructing a robust protein-protein interaction network by integrating multiple public databases. *BMC Bioinformatics*, *12 Suppl 10*, S7. <https://doi.org/10.1186/1471-2105-12-S10-S7>
- Martínez-García, A., Alvarez-Romero, C., Román-Villarán, E., Bernabeu-Wittel, M., & Luis Parra-Calderón, C. (2023). FAIR principles to improve the impact on health research management outcomes. *Heliyon*, *9*(5), e15733. <https://doi.org/10.1016/j.heliyon.2023.e15733>
- Marwah, V. S., Kinaret, P. A. S., Serra, A., Scala, G., Lauerma, A., Fortino, V., & Greco, D. (2018). Inform: inference of network response modules. *Bioinformatics*, *34*(12), 2136–2138. <https://doi.org/10.1093/bioinformatics/bty063>
- Marx, V. (2013). Biology: The big challenges of big data. *Nature*, *498*(7453), 255–260. <https://doi.org/10.1038/498255a>
- Mayo, C. S., Matuszak, M. M., Schipper, M. J., Jolly, S., Hayman, J. A., & Ten Haken, R. K. (2017). Big data in designing clinical trials: opportunities and challenges. *Frontiers in Oncology*, *7*, 187. <https://doi.org/10.3389/fonc.2017.00187>
- Mehta, N., & Pandit, A. (2018). Concurrence of big data analytics and healthcare: A systematic review. *International Journal of Medical Informatics*, *114*, 57–65. <https://doi.org/10.1016/j.ijmedinf.2018.03.013>
- Meng, L., & Xiang, J. (2018). Brain network analysis and classification based on convolutional neural network. *Frontiers in Computational Neuroscience*, *12*, 95. <https://doi.org/10.3389/fncom.2018.00095>
- Merkel, D. (2014). Docker: Lightweight Linux Containers for Consistent Development and Deployment. *Linux J.*, *2014*(239).

- Merwar, G., Gibbons, J. R., Hosseini, S. A., & Saadabadi, A. (2023). Nortriptyline. In *StatPearls*. StatPearls Publishing.
- Meyers, J., Fabian, B., & Brown, N. (2021). De novo molecular design and generative models. *Drug Discovery Today*, 26(11), 2707–2715.  
<https://doi.org/10.1016/j.drudis.2021.05.019>
- Meyer, P. E., Kontos, K., Lafitte, F., & Bontempi, G. (2007). Information-theoretic inference of large transcriptional regulatory networks. *EURASIP Journal on Bioinformatics & Systems Biology*, 79879. <https://doi.org/10.1155/2007/79879>
- Meyer, P. E., Lafitte, F., & Bontempi, G. (2008). minet: A R/Bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinformatics*, 9, 461. <https://doi.org/10.1186/1471-2105-9-461>
- Mi, H., Ebert, D., Muruganujan, A., Mills, C., Albu, L.-P., Mushayamaha, T., & Thomas, P. D. (2021). PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Research*, 49(D1), D394–D403. <https://doi.org/10.1093/nar/gkaa1106>
- Mihaylov, I., Kañdula, M., Krachunov, M., & Vassilev, D. (2019). A novel framework for horizontal and vertical data integration in cancer studies with application to survival time prediction models. *Biology Direct*, 14(1), 22.  
<https://doi.org/10.1186/s13062-019-0249-6>
- Mihaylov, I., Nisheva-Pavlova, M., & Vassilev, D. (2019). An approach for semantic data integration in cancer studies. In J. M. F. Rodrigues, P. J. S. Cardoso, J. Monteiro, R. Lam, V. V. Krzhizhanovskaya, M. H. Lees, J. J. Dongarra, & P. M. A. Sloot (Eds.), *Computational science – ICCS 2019: 19th international conference, faro, portugal, june 12–14, 2019, proceedings, part III* (Vol. 11538, pp. 60–73). Springer International Publishing. [https://doi.org/10.1007/978-3-030-22744-9\\_5](https://doi.org/10.1007/978-3-030-22744-9_5)

- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *ArXiv*. <https://doi.org/10.48550/arxiv.1301.3781>
- Miller, J. C., Skoll, D., & Saxon, L. A. (2020). Home Monitoring of Cardiac Devices in the Era of COVID-19. *Current Cardiology Reports*, 23(1), 1. <https://doi.org/10.1007/s11886-020-01431-w>
- Mohamed, S. K., Nounu, A., & Nováček, V. (2019). Drug target discovery using knowledge graph embeddings. *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing - SAC '19*, 11–18. <https://doi.org/10.1145/3297280.3297282>
- Mohamed, S. K., Nováček, V., & Nounu, A. (2020). Discovering protein drug targets using knowledge graph embeddings. *Bioinformatics*, 36(2), 603–610. <https://doi.org/10.1093/bioinformatics/btz600>
- Mouchlis, V. D., Afantitis, A., Serra, A., Fratello, M., Papadiamantis, A. G., Aidinis, V., Lynch, I., Greco, D., & Melagraki, G. (2021). Advances in de novo drug design: from conventional to machine learning methods. *International Journal of Molecular Sciences*, 22(4). <https://doi.org/10.3390/ijms22041676>
- Mungall, C. J., McMurry, J. A., Köhler, S., Balhoff, J. P., Borromeo, C., Brush, M., Carbon, S., Conlin, T., Dunn, N., Engelstad, M., Foster, E., Gourdine, J. P., Jacobsen, J. O. B., Keith, D., Laraway, B., Lewis, S. E., NguyenXuan, J., Shefchek, K., Vasilevsky, N., ... Haendel, M. A. (2017). The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Research*, 45(D1), D712–D722. <https://doi.org/10.1093/nar/gkw1128>
- Myklebust, E. B., Jimenez-Ruiz, E., Chen, J., Wolf, R., & Tollefsen, K. E. (n.d.). Prediction of Adverse Biological Effects of Chemicals Using Knowledge Graph Embeddings. *Semantic Web – Interoperability, Usability, Applicability*.

- Naima, R., Imen, M., Mustapha, J., Hafedh, A., Kamel, K., Mohsen, S., & Salem, A. (2021). Acute titanium dioxide nanoparticles exposure impaired spatial cognitive performance through neurotoxic and oxidative mechanisms in Wistar rats. *Biomarkers: Biochemical Indicators of Exposure, Response, and Susceptibility To Chemicals*, 26(8), 760–769. <https://doi.org/10.1080/1354750X.2021.1999501>
- National Center for Biotechnology Information. (n.d.). *PubChem Compound Summary for CID 4641, Oxyphenbutazone*. . National Center for Biotechnology Information. Retrieved March 22, 2023, from <https://pubchem.ncbi.nlm.nih.gov/compound/Oxyphenbutazone>
- Nelson, M. R., Johnson, T., Warren, L., Hughes, A. R., Chissoe, S. L., Xu, C.-F., & Waterworth, D. M. (2016). The genetics of drug efficacy: opportunities and challenges. *Nature Reviews. Genetics*, 17(4), 197–206. <https://doi.org/10.1038/nrg.2016.12>
- Nelson, S. J., Zeng, K., Kilbourne, J., Powell, T., & Moore, R. (2011). Normalized names for clinical drugs: RxNorm at 6 years. *Journal of the American Medical Informatics Association*, 18(4), 441–448. <https://doi.org/10.1136/amiajnl-2011-000116>
- Newman, M. E. J. (2005). A measure of betweenness centrality based on random walks. *Social Networks*, 27(1), 39–54. <https://doi.org/10.1016/j.socnet.2004.11.009>
- Nguyen, D. A., Nguyen, C. H., & Mamitsuka, H. (2021). A survey on adverse drug reaction studies: data, tasks and machine learning methods. *Briefings in Bioinformatics*, 22(1), 164–177. <https://doi.org/10.1093/bib/bbz140>
- Nickel, M., Murphy, K., Tresp, V., & Gabrilovich, E. (2016). A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1), 11–33. <https://doi.org/10.1109/JPROC.2015.2483592>

- Nimpf, S., & Keays, D. A. (2020). Why (and how) we should publish negative data. *EMBO Reports*, 21(1), e49775. <https://doi.org/10.15252/embr.201949775>
- Nováček, V., & Mohamed, S. K. (2020). Predicting Polypharmacy Side-effects Using Knowledge Graph Embeddings. *AMLA Joint Summits on Translational Science Proceedings AMLA Summit on Translational Science, 2020*, 449–458.
- Ochoa, D., Hercules, A., Carmona, M., Suveges, D., Baker, J., Malangone, C., Lopez, I., Miranda, A., Cruz-Castillo, C., Fumis, L., Bernal-Llinares, M., Tsukanov, K., Cornu, H., Tsirigos, K., Razuvayevskaya, O., Buniello, A., Schwartzentruber, J., Karim, M., Ariano, B., ... McDonagh, E. M. (2023). The next-generation Open Targets Platform: reimagined, redesigned, rebuilt. *Nucleic Acids Research*, 51(D1), D1353–D1359. <https://doi.org/10.1093/nar/gkac1046>
- Odibat, O., & Reddy, C. K. (2012). Ranking differential hubs in gene co-expression networks. *Journal of Bioinformatics and Computational Biology*, 10(1), 1240002. <https://doi.org/10.1142/S0219720012400021>
- Oestreich, M., Holsten, L., Agrawal, S., Dahm, K., Koch, P., Jin, H., Becker, M., & Ulas, T. (2022). hCoCena: horizontal integration and analysis of transcriptomics datasets. *Bioinformatics*, 38(20), 4727–4734. <https://doi.org/10.1093/bioinformatics/btac589>
- Ovens, K., Eames, B. F., & McQuillan, I. (2021). Comparative Analyses of Gene Co-expression Networks: Implementations and Applications in the Study of Evolution. *Frontiers in Genetics*, 12, 695399. <https://doi.org/10.3389/fgene.2021.695399>
- Pancino, N., Perron, Y., Bongini, P., & Scarselli, F. (2022). Drug Side Effect Prediction with Deep Learning Molecular Embedding in a Graph-of-Graphs Domain. *Mathematics*, 10(23), 4550. <https://doi.org/10.3390/math10234550>

- Papatheodorou, I., Fonseca, N. A., Keays, M., Tang, Y. A., Barrera, E., Bazant, W., Burke, M., Füllgrabe, A., Fuentes, A. M.-P., George, N., Huerta, L., Koskinen, S., Mohammed, S., Geniza, M., Preece, J., Jaiswal, P., Jarnuczak, A. F., Huber, W., Stegle, O., ... Petryszak, R. (2018). Expression Atlas: gene and protein expression across multiple studies and organisms. *Nucleic Acids Research*, *46*(D1), D246–D251. <https://doi.org/10.1093/nar/gkx1158>
- Patil, A., Nakai, K., & Nakamura, H. (2011). HitPredict: a database of quality assessed protein-protein interactions in nine species. *Nucleic Acids Research*, *39*(Database issue), D744-9. <https://doi.org/10.1093/nar/gkq897>
- Paton, K. (1969). An algorithm for finding a fundamental set of cycles of a graph. *Communications of the ACM*, *12*(9), 514–518. <https://doi.org/10.1145/363219.363232>
- Pavel, A., Saarimäki, L. A., Möbus, L., Federico, A., Serra, A., & Greco, D. (2022). The potential of a data centred approach & knowledge graph data representation in chemical safety and drug design. *Computational and Structural Biotechnology Journal*, *20*, 4837–4849. <https://doi.org/10.1016/j.csbj.2022.08.061>
- Pavel, A., Serra, A., Cattelani, L., Federico, A., & Greco, D. (2022). Network analysis of microarray data. *Methods in Molecular Biology*, *2401*, 161–186. [https://doi.org/10.1007/978-1-0716-1839-4\\_11](https://doi.org/10.1007/978-1-0716-1839-4_11)
- Paysan-Lafosse, T., Blum, M., Chuguransky, S., Grego, T., Pinto, B. L., Salazar, G. A., Bileschi, M. L., Bork, P., Bridge, A., Colwell, L., Gough, J., Haft, D. H., Letunić, I., Marchler-Bauer, A., Mi, H., Natale, D. A., Orengo, C. A., Pandurangan, A. P., Rivoire, C., ... Bateman, A. (2023). InterPro in 2022. *Nucleic Acids Research*, *51*(D1), D418–D427. <https://doi.org/10.1093/nar/gkac993>

- Phung, K. A., Nguyen, T. T., Wangad, N., Baraheem, S., Vo, N. D., & Nguyen, K. (2022). Disease Recognition in X-ray Images with Doctor Consultation-Inspired Model. *Journal of Imaging*, 8(12). <https://doi.org/10.3390/jimaging8120323>
- Pierson, E., GTEx Consortium, Koller, D., Battle, A., Mostafavi, S., Ardlie, K. G., Getz, G., Wright, F. A., Kellis, M., Volpi, S., & Dermitzakis, E. T. (2015). Sharing and Specificity of Co-expression Networks across 35 Human Tissues. *PLoS Computational Biology*, 11(5), e1004220. <https://doi.org/10.1371/journal.pcbi.1004220>
- Piñero, J., Ramírez-Angueta, J. M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., & Furlong, L. I. (2020). The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research*, 48(D1), D845–D855. <https://doi.org/10.1093/nar/gkz1021>
- Prashanth, G., Vastrad, B., Tengli, A., Vastrad, C., & Kotturshetti, I. (2021). Identification of hub genes related to the progression of type 1 diabetes by computational analysis. *BMC Endocrine Disorders*, 21(1), 61. <https://doi.org/10.1186/s12902-021-00709-6>
- Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A., & Murali, T. M. (2020). Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature Methods*, 17(2), 147–154. <https://doi.org/10.1038/s41592-019-0690-6>
- Preto, A. J., Correia, P. C., & Moreira, I. S. (2022). DrugTax: package for drug taxonomy identification and explainable feature extraction. *Journal of Cheminformatics*, 14(1), 73. <https://doi.org/10.1186/s13321-022-00649-w>
- Proost, S., & Mutwil, M. (2018). CoNekT: an open-source framework for comparative genomic and transcriptomic network analyses. *Nucleic Acids Research*, 46(W1), W133–W140. <https://doi.org/10.1093/nar/gky336>

- Przulj, N., Corneil, D. G., & Jurisica, I. (2006). Efficient estimation of graphlet frequency distributions in protein-protein interaction networks. *Bioinformatics*, 22(8), 974–980. <https://doi.org/10.1093/bioinformatics/btl030>
- Przulj, N. (2007). Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2), e177-83. <https://doi.org/10.1093/bioinformatics/btl301>
- Qian, T., Zhu, S., & Hoshida, Y. (2019). Use of big data in drug development for precision medicine: an update. *Expert Review of Precision Medicine and Drug Development*, 4(3), 189–200. <https://doi.org/10.1080/23808993.2019.1617632>
- Queralt-Rosinach, N., Kaliyaperumal, R., Bernabé, C. H., Long, Q., Joosten, S. A., van der Wijk, H. J., Flikkenschild, E. L. A., Burger, K., Jacobsen, A., Mons, B., Roos, M., BEAT-COVID Group, & COVID-19 LUMC Group. (2022). Applying the FAIR principles to data in a hospital: challenges and opportunities in a pandemic. *Journal of Biomedical Semantics*, 13(1), 12. <https://doi.org/10.1186/s13326-022-00263-7>
- Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., Liu, P. J., Liu, X., Marcus, J., Sun, M., Sundberg, P., Yee, H., Zhang, K., Zhang, Y., Flores, G., Duggan, G. E., Irvine, J., Le, Q., Litsch, K., ... Dean, J. (2018). Scalable and accurate deep learning with electronic health records. *Npj Digital Medicine*, 1, 18. <https://doi.org/10.1038/s41746-018-0029-1>
- Ramamoorthy, A., Pacanowski, M. A., Bull, J., & Zhang, L. (2015). Racial/ethnic differences in drug disposition and response: review of recently approved drugs. *Clinical Pharmacology and Therapeutics*, 97(3), 263–273. <https://doi.org/10.1002/cpt.61>
- Rao, M. S., Van Vleet, T. R., Ciurlionis, R., Buck, W. R., Mittelstadt, S. W., Blomme, E. A. G., & Liguori, M. J. (2018). Comparison of RNA-Seq and Microarray Gene Expression Platforms for the Toxicogenomic Evaluation of Liver From Short-Term



- Rat Toxicity Studies. *Frontiers in Genetics*, *9*, 636.  
<https://doi.org/10.3389/fgene.2018.00636>
- Ratajczak, F., Joblin, M., Ringsquandl, M., & Hildebrandt, M. (2022). Task-driven knowledge graph filtering improves prioritizing drugs for repurposing. *BMC Bioinformatics*, *23*(1), 84. <https://doi.org/10.1186/s12859-022-04608-y>
- Ravera, F., Cirmena, G., Dameri, M., Gallo, M., Vellone, V. G., Fregatti, P., Friedman, D., Calabrese, M., Ballestrero, A., Tagliafico, A., Ferrando, L., & Zoppoli, G. (2021). Development of a hoRizontal data intEgration classifier for NOn-invasive early diAgnosis of breasT cancEr: the RENOVATE study protocol. *BMJ Open*, *11*(12), e054256. <https://doi.org/10.1136/bmjopen-2021-054256>
- Razaghi-Moghadam, Z., & Nikoloski, Z. (2020). Supervised learning of gene-regulatory networks based on graph distance profiles of transcriptomics data. *NPJ Systems Biology and Applications*, *6*(1), 21. <https://doi.org/10.1038/s41540-020-0140-1>
- Ren, Y., Ay, A., & Kahveci, T. (2018). Shortest path counting in probabilistic biological networks. *BMC Bioinformatics*, *19*(1), 465. <https://doi.org/10.1186/s12859-018-2480-z>
- Richard, A. M., Huang, R., Waidyanatha, S., Shinn, P., Collins, B. J., Thillainadarajah, I., Grulke, C. M., Williams, A. J., Lougee, R. R., Judson, R. S., Houck, K. A., Shobair, M., Yang, C., Rathman, J. F., Yasgar, A., Fitzpatrick, S. C., Simeonov, A., Thomas, R. S., Crofton, K. M., ... Tice, R. R. (2021). The tox21 10K compound library: collaborative chemistry advancing toxicology. *Chemical Research in Toxicology*, *34*(2), 189–216. <https://doi.org/10.1021/acs.chemrestox.0c00264>
- Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and

- microarray studies. *Nucleic Acids Research*, *43*(7), e47.  
<https://doi.org/10.1093/nar/gkv007>
- Rolland, T., Taşan, M., Charloteaux, B., Pevzner, S. J., Zhong, Q., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., Mosca, R., Kamburov, A., Ghiassian, S. D., Yang, X., Ghamsari, L., Balcha, D., Begg, B. E., Braun, P., Brehme, M., Broly, M. P., ... Vidal, M. (2014). A proteome-scale map of the human interactome network. *Cell*, *159*(5), 1212–1226. <https://doi.org/10.1016/j.cell.2014.10.050>
- Rossato, L. G., Costa, V. M., Dallegrave, E., Arbo, M., Silva, R., Ferreira, R., Amado, F., Dinis-Oliveira, R. J., Duarte, J. A., de Lourdes Bastos, M., Palmeira, C., & Remião, F. (2014). Mitochondrial cumulative damage induced by mitoxantrone: late onset cardiac energetic impairment. *Cardiovascular Toxicology*, *14*(1), 30–40.  
<https://doi.org/10.1007/s12012-013-9230-2>
- Rosvall, M., & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(4), 1118–1123. <https://doi.org/10.1073/pnas.0706851105>
- Ruiz, P., Beglitti, G., Tincher, T., Wheeler, J., & Mumtaz, M. (2012). Prediction of acute mammalian toxicity using QSAR methods: a case study of sulfur mustard and its breakdown products. *Molecules (Basel, Switzerland)*, *17*(8), 8982–9001.  
<https://doi.org/10.3390/molecules17088982>
- Rukov, J. L., Wilentzik, R., Jaffe, I., Vinther, J., & Shomron, N. (2014). Pharmaco-miR: linking microRNAs and drug effects. *Briefings in Bioinformatics*, *15*(4), 648–659.  
<https://doi.org/10.1093/bib/bbs082>
- Saarimäki, L. A., Melagraki, G., Afantitis, A., Lynch, I., & Greco, D. (2022). Prospects and challenges for FAIR toxicogenomics data. *Nature Nanotechnology*, *17*(1), 17–18.  
<https://doi.org/10.1038/s41565-021-01049-1>

- Saarimäki, L. A., Kinaret, P. A. S., Scala, G., del Giudice, G., Federico, A., Serra, A., & Greco, D. (2020). Toxicogenomics analysis of dynamic dose-response in macrophages highlights molecular alterations relevant for multi-walled carbon nanotube-induced lung fibrosis. *NanoImpact*, 100274.  
<https://doi.org/10.1016/j.impact.2020.100274>
- Saarimäki, L., Federico, A., Lynch, I., Papadiamantis, A. G., Tsoumanis, A., Melagraki, G., Afantitis, A., Serra, A., & Greco, D. (2021). Manually curated transcriptomics data collection for toxicogenomic assessment of engineered nanomaterials. *Scientific Data*, 8(1), 49. <https://doi.org/10.1038/s41597-021-00808-y>
- Saarimäki, L., Fratello, M., Pavel, A., Korpilähde, S., Leppänen, J., Serra, A., & Greco, D. (2023). A curated gene and biological system annotation of adverse outcome pathways related to human health. *Scientific Data*, 10(1), 409.  
<https://doi.org/10.1038/s41597-023-02321-w>
- Saarimäki, L., Morikka, J., Pavel, A., Korpilähde, S., Del Giudice, G., Federico, A., Fratello, M., Serra, A., & Greco, D. (2023). Toxicogenomics Data for Chemical Safety Assessment and Development of New Approach Methodologies: An Adverse Outcome Pathway-Based Approach. *Advanced Science (Weinheim, Baden-Württemberg, Germany)*, 10(2), e2203984. <https://doi.org/10.1002/advs.202203984>
- Sagioglu, S., & Sinanc, D. (2013). Big data: A review. *2013 International Conference on Collaboration Technologies and Systems (CTS)*, 42–47.  
<https://doi.org/10.1109/CTS.2013.6567202>
- Salvatore, C., Cerasa, A., Castiglioni, I., Gallivanone, F., Augimeri, A., Lopez, M., Arabia, G., Morelli, M., Gilardi, M. C., & Quattrone, A. (2014). Machine learning on brain MRI data for differential diagnosis of Parkinson's disease and Progressive

- Supranuclear Palsy. *Journal of Neuroscience Methods*, 222, 230–237.  
<https://doi.org/10.1016/j.jneumeth.2013.11.016>
- Sandelin, A., Alkema, W., Engström, P., Wasserman, W. W., & Lenhard, B. (2004). JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research*, 32(Database issue), D91-4.  
<https://doi.org/10.1093/nar/gkh012>
- Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., Connor, R., Funk, K., Kelly, C., Kim, S., Madej, T., Marchler-Bauer, A., Lanczycki, C., Lathrop, S., Lu, Z., Thibaud-Nissen, F., Murphy, T., Phan, L., Skripchenko, Y., ... Sherry, S. T. (2022). Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 50(D1), D20–D26.  
<https://doi.org/10.1093/nar/gkab1112>
- Scheuermann, R. H., Ceusters, W., & Smith, B. (2009). Toward an ontological treatment of disease and diagnosis. *Summit on Translational Bioinformatics, 2009*, 116–120.
- Schoch, C. L., Ciufo, S., Domrachev, M., Hotton, C. L., Kannan, S., Khovanskaya, R., Leipe, D., Mcveigh, R., O'Neill, K., Robbertse, B., Sharma, S., Soussov, V., Sullivan, J. P., Sun, L., Turner, S., & Karsch-Mizrachi, I. (2020). NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database: The Journal of Biological Databases and Curation*, 2020. <https://doi.org/10.1093/database/baaa062>
- Schriml, L. M., Lichenstein, R., Bisordi, K., Bearer, C., Baron, J. A., & Greene, C. (2023). Modeling the enigma of complex disease etiology. *Journal of Translational Medicine*, 21(1), 148. <https://doi.org/10.1186/s12967-023-03987-x>

- Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and Statistical Modeling with Python. *Proceedings of the 9th Python in Science Conference*, 92–96.  
<https://doi.org/10.25080/Majora-92bf1922-011>
- Serra, A., Fratello, M., Cattelani, L., Liampa, I., Melagraki, G., Kohonen, P., Nymark, P., Federico, A., Kinaret, P. A. S., Jagiello, K., Ha, M. K., Choi, J.-S., Sanabria, N., Gulumian, M., Puzyn, T., Yoon, T.-H., Sarimveis, H., Grafström, R., Afantitis, A., & Greco, D. (2020). Transcriptomics in toxicogenomics, part III: data modelling for risk assessment. *Nanomaterials (Basel, Switzerland)*, 10(4).  
<https://doi.org/10.3390/nano10040708>
- Serra, A., Fratello, M., Federico, A., Ojha, R., Provenzani, R., Tasnadi, E., Cattelani, L., Del Giudice, G., Kinaret, P. A. S., Saarimäki, L. A., Pavel, A., Kuivanen, S., Cerullo, V., Vapalahti, O., Horvath, P., Lieto, A. D., Yli-Kauhaluoma, J., Balistreri, G., & Greco, D. (2022). Computationally prioritized drugs inhibit SARS-CoV-2 infection and syncytia formation. *Briefings in Bioinformatics*, 23(1).  
<https://doi.org/10.1093/bib/bbab507>
- Serra, A., Fratello, M., Fortino, V., Raiconi, G., Tagliaferri, R., & Greco, D. (2015). MVDA: a multi-view genomic data integration methodology. *BMC Bioinformatics*, 16, 261. <https://doi.org/10.1186/s12859-015-0680-3>
- Serra, A., Letunic, I., Fortino, V., Handy, R. D., Fadeel, B., Tagliaferri, R., & Greco, D. (2019). INSIdE NANO: a systems biology framework to contextualize the mechanism-of-action of engineered nanomaterials. *Scientific Reports*, 9(1), 179.  
<https://doi.org/10.1038/s41598-018-37411-y>
- Serra, A., Saarimäki, L. A., Pavel, A., Del Giudice, G., Fratello, M., Cattelani, L., Federico, A., Laurino, O., Marwah, V. S., Fortino, V., Scala, G., Sofia Kinaret, P. A., & Greco, D. (2022). Nextcast: A software suite to analyse and model

- toxicogenomics data. *Computational and Structural Biotechnology Journal*, 20, 1413–1426.  
<https://doi.org/10.1016/j.csbj.2022.03.014>
- Sezer Tuncsoy, B., Tuncsoy, M., Gomes, T., Sousa, V., Teixeira, M. R., Bebianno, M. J., & Ozalp, P. (2019). Effects of Copper Oxide Nanoparticles on Tissue Accumulation and Antioxidant Enzymes of *Galleria mellonella* L. *Bulletin of Environmental Contamination and Toxicology*, 102(3), 341–346.  
<https://doi.org/10.1007/s00128-018-2529-8>
- Sharifi, S., Caracciolo, G., Pozzi, D., Digiacomio, L., Swann, J., Daldrup-Link, H. E., & Mahmoudi, M. (2021). The role of sex as a biological variable in the efficacy and toxicity of therapeutic nanomedicine. *Advanced Drug Delivery Reviews*, 174, 337–347.  
<https://doi.org/10.1016/j.addr.2021.04.028>
- Sharma, B., Chenthamarakshan, V., Dhurandhar, A., Pereira, S., Hendler, J. A., Dordick, J. S., & Das, P. (2023). Accurate clinical toxicity prediction using multi-task deep neural nets and contrastive molecular explanations. *Scientific Reports*, 13(1), 4908. <https://doi.org/10.1038/s41598-023-31169-8>
- Sheth, A., Padhee, S., Gyrard, A., & Sheth, A. (2019). Knowledge graphs and knowledge networks: the story in brief. *IEEE Internet Computing*, 23(4), 67–75.  
<https://doi.org/10.1109/MIC.2019.2928449>
- Silva, M. C., Eugénio, P., Faria, D., & Pesquita, C. (2022). Ontologies and knowledge graphs in oncology research. *Cancers*, 14(8).  
<https://doi.org/10.3390/cancers14081906>
- Simões, S. N., Martins-Jr, D. C., Brentani, H., & Fumio, R. (2012). Shortest paths ranking methodology to identify alterations in PPI networks of complex diseases. *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine - BCB '12*, 561–563. <https://doi.org/10.1145/2382936.2383021>

- Smith, C. L., & Eppig, J. T. (2009). The mammalian phenotype ontology: enabling robust annotation and comparative analysis. *Wiley Interdisciplinary Reviews. Systems Biology and Medicine*, 1(3), 390–399. <https://doi.org/10.1002/wsbm.44>
- Sollis, E., Mosaku, A., Abid, A., Buniello, A., Cerezo, M., Gil, L., Groza, T., Güneş, O., Hall, P., Hayhurst, J., Ibrahim, A., Ji, Y., John, S., Lewis, E., MacArthur, J. A. L., McMahon, A., Osumi-Sutherland, D., Panoutsopoulou, K., Pendlington, Z., ... Harris, L. W. (2023). The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Research*, 51(D1), D977–D985. <https://doi.org/10.1093/nar/gkac1010>
- Song, Z.-Y., Chao, F., Zhuo, Z., Ma, Z., Li, W., & Chen, G. (2019). Identification of hub genes in prostate cancer using robust rank aggregation and weighted gene co-expression network analysis. *Aging*, 11(13), 4736–4756. <https://doi.org/10.18632/aging.102087>
- Sonmez, A. B., & Can, T. (2017). Comparison of tissue/disease specific integrated networks using directed graphlet signatures. *BMC Bioinformatics*, 18(Suppl 4), 135. <https://doi.org/10.1186/s12859-017-1525-z>
- Sosa, D. N., Derry, A., Guo, M., Wei, E., Brinton, C., & Altman, R. B. (2020). A Literature-Based Knowledge Graph Embedding Method for Identifying Drug Repurposing Opportunities in Rare Diseases. *Pacific Symposium on Biocomputing*, 25, 463–474.
- Steenwinckel, B., Vandewiele, G., Weyns, M., Agozzino, T., Turck, F. D., & Ongenaes, F. (2022). INK: knowledge graph embeddings for node classification. *Data Mining and Knowledge Discovery*, 36(2), 620–667. <https://doi.org/10.1007/s10618-021-00806-z>

- Subrahmanya, S. V. G., Shetty, D. K., Patil, V., Hameed, B. M. Z., Paul, R., Smriti, K., Naik, N., & Somani, B. K. (2022). The role of data science in healthcare advancements: applications, benefits, and future prospects. *Irish Journal of Medical Science*, 191(4), 1473–1483. <https://doi.org/10.1007/s11845-021-02730-z>
- Subramanian, A., Narayan, R., Corsello, S. M., Peck, D. D., Natoli, T. E., Lu, X., Gould, J., Davis, J. F., Tubelli, A. A., Asiedu, J. K., Lahr, D. L., Hirschman, J. E., Liu, Z., Donahue, M., Julian, B., Khan, M., Wadden, D., Smith, I. C., Lam, D., ... Golub, T. R. (2017). A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6), 1437-1452.e17. <https://doi.org/10.1016/j.cell.2017.10.049>
- Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43), 15545–15550. <https://doi.org/10.1073/pnas.0506580102>
- Sun, T., Kang, Y., Liu, J., Zhang, Y., Ou, L., Liu, X., Lai, R., & Shao, L. (2021). Nanomaterials and hepatic disease: toxicokinetics, disease types, intrinsic mechanisms, liver susceptibility, and influencing factors. *Journal of Nanobiotechnology*, 19(1), 108. <https://doi.org/10.1186/s12951-021-00843-2>
- Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K. P., Kuhn, M., Bork, P., Jensen, L. J., & von Mering, C. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, 43(Database issue), D447-52. <https://doi.org/10.1093/nar/gku1003>



- Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N. T., Morris, J. H., Bork, P., Jensen, L. J., & Mering, C. von. (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, *47*(D1), D607–D613. <https://doi.org/10.1093/nar/gky1131>
- Szklarczyk, D., Santos, A., von Mering, C., Jensen, L. J., Bork, P., & Kuhn, M. (2016). STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data. *Nucleic Acids Research*, *44*(D1), D380-4. <https://doi.org/10.1093/nar/gkv1277>
- Tabula Sapiens Consortium\*, Jones, R. C., Karkanias, J., Krasnow, M. A., Pisco, A. O., Quake, S. R., Salzman, J., Yosef, N., Bulthaupt, B., Brown, P., Harper, W., Hemenez, M., Ponnusamy, R., Salehi, A., Sanagavarapu, B. A., Spallino, E., Aaron, K. A., Concepcion, W., Gardner, J. M., ... et al. (2022). The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science*, *376*(6594), eabl4896. <https://doi.org/10.1126/science.abl4896>
- Tandon, A., Albeshri, A., Thayanathan, V., Alhalabi, W., & Fortunato, S. (2019). Fast consensus clustering in complex networks. *Physical Review. E*, *99*(4–1), 042301. <https://doi.org/10.1103/PhysRevE.99.042301>
- Tang, D., Zhao, X., Zhang, L., Wang, Z., & Wang, C. (2019). Identification of hub genes to regulate breast cancer metastasis to brain by bioinformatics analyses. *Journal of Cellular Biochemistry*, *120*(6), 9522–9531. <https://doi.org/10.1002/jcb.28228>
- Tanvir, R. B., & Mondal, A. M. (2019). Cancer Biomarker Discovery from Gene Co-expression Networks Using Community Detection Methods. *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2097–2104. <https://doi.org/10.1109/BIBM47256.2019.8982960>

- Taramasco, C., Olivares, R., Munoz, R., Soto, R., Villar, M., & de Albuquerque, V. H. C. (2019). The patient bed assignment problem solved by autonomous bat algorithm. *Applied Soft Computing*, *81*, 105484. <https://doi.org/10.1016/j.asoc.2019.105484>
- Thafar, M. A., Olayan, R. S., Ashoor, H., Albaradei, S., Bajic, V. B., Gao, X., Gojobori, T., & Essack, M. (2020). DTiGEMS+: drug-target interaction prediction using graph embedding, graph mining, and similarity-based techniques. *Journal of Cheminformatics*, *12*(1), 44. <https://doi.org/10.1186/s13321-020-00447-2>
- The Gene Ontology Consortium. (2021). The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Research*, *49*(D1), D325–D334. <https://doi.org/10.1093/nar/gkaa1113>
- Thioridazine. (2012). In *LiverTox: Clinical and Research Information on Drug-Induced Liver Injury*. National Institute of Diabetes and Digestive and Kidney Diseases.
- Thul, P. J., & Lindskog, C. (2018). The human protein atlas: A spatial map of the human proteome. *Protein Science*, *27*(1), 233–244. <https://doi.org/10.1002/pro.3307>
- Tiddi, I., Balliet, D., & ten Teije, A. (2020). Fostering Scientific Meta-analyses with Knowledge Graphs: A Case-Study. In A. Harth, S. Kirrane, A.-C. Ngonga Ngomo, H. Paulheim, A. Rula, A. L. Gentile, P. Haase, & M. Cochez (Eds.), *The semantic web: 17th international conference, ESWC 2020, heraklion, crete, greece, may 31–june 4, 2020, proceedings* (Vol. 12123, pp. 287–303). Springer International Publishing. [https://doi.org/10.1007/978-3-030-49461-2\\_17](https://doi.org/10.1007/978-3-030-49461-2_17)
- Tolani, P., Gupta, S., Yadav, K., Aggarwal, S., & Yadav, A. K. (2021). Big data, integrative omics and network biology. *Advances in Protein Chemistry and Structural Biology*, *127*, 127–160. <https://doi.org/10.1016/bs.apcsb.2021.03.006>

- Tong, X., Liu, X., Tan, X., Li, X., Jiang, J., Xiong, Z., Xu, T., Jiang, H., Qiao, N., & Zheng, M. (2021). Generative models for de novo drug design. *Journal of Medicinal Chemistry*, *64*(19), 14011–14027. <https://doi.org/10.1021/acs.jmedchem.1c00927>
- Tong, Z., Cui, Q., Wang, J., & Zhou, Y. (2019). TransmiR v2.0: an updated transcription factor-microRNA regulation database. *Nucleic Acids Research*, *47*(D1), D253–D258. <https://doi.org/10.1093/nar/gky1023>
- Tu, D., Ma, C., Zeng, Z., Xu, Q., Guo, Z., Song, X., & Zhao, X. (2022). Identification of hub genes and transcription factor regulatory network for heart failure using RNA-seq data and robust rank aggregation analysis. *Frontiers in Cardiovascular Medicine*, *9*, 916429. <https://doi.org/10.3389/fcvm.2022.916429>
- Ulfenborg, B. (2019). Vertical and horizontal integration of multi-omics data with miodin. *BMC Bioinformatics*, *20*(1), 649. <https://doi.org/10.1186/s12859-019-3224-4>
- Unsal-Beyge, S., & Tuncbag, N. (2022). Functional stratification of cancer drugs through integrated network similarity. *NPJ Systems Biology and Applications*, *8*(1), 11. <https://doi.org/10.1038/s41540-022-00219-8>
- Urbanski, A. H., Araujo, J. D., Creighton, R., & Nakaya, H. I. (2019). Integrative biology approaches applied to human diseases. In H. Husi (Ed.), *Computational Biology*. Codon Publications. <https://doi.org/10.15586/computationalbiology.2019.ch2>
- Vashishth, S., Sanyal, S., Nitin, V., Agrawal, N., & Talukdar, P. (2020). InteractE: Improving Convolution-Based Knowledge Graph Embeddings by Increasing Feature Interactions. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(03), 3009–3016. <https://doi.org/10.1609/aaai.v34i03.5694>
- Veena, D. K., Jatti, A., Joshi, R., & Deepu, K. S. (2017). Characterization of dental pathologies using digital panoramic X-ray images based on texture analysis. *Annual*

- International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference, 2017*, 592–595.  
<https://doi.org/10.1109/EMBC.2017.8036894>
- Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., ... SciPy 1.0 Contributors. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3), 261–272.  
<https://doi.org/10.1038/s41592-019-0686-2>
- Vogenberg, F. R., Isaacson Barash, C., & Pursel, M. (2010). Personalized medicine: part 1: evolution and development into theranostics. *P & T: A Peer-Reviewed Journal for Formulary Management*, 35(10), 560–576.
- Wang, F., Zhang, P., Cao, N., Hu, J., & Sorrentino, R. (2014). Exploring the associations between drug side-effects and therapeutic indications. *Journal of Biomedical Informatics*, 51, 15–23. <https://doi.org/10.1016/j.jbi.2014.03.014>
- Wang, J., Lu, M., Qiu, C., & Cui, Q. (2010). TransmiR: a transcription factor-microRNA regulation database. *Nucleic Acids Research*, 38(Database issue), D119-22.  
<https://doi.org/10.1093/nar/gkp803>
- Wang, Meng, Wang, H., Liu, X., Ma, X., & Wang, B. (2021). Drug-Drug Interaction Predictions via Knowledge Graph and Text Embedding: Instrument Validation Study. *JMIR Medical Informatics*, 9(6), e28277. <https://doi.org/10.2196/28277>
- Wang, Mingyang, Wang, Z., Sun, H., Wang, J., Shen, C., Weng, G., Chai, X., Li, H., Cao, D., & Hou, T. (2022). Deep learning approaches for de novo drug design: An overview. *Current Opinion in Structural Biology*, 72, 135–144.  
<https://doi.org/10.1016/j.sbi.2021.10.001>

- Wang, Shudong, Du, Z., Ding, M., Rodriguez-Paton, A., & Song, T. (2022). KG-DTI: a knowledge graph based deep learning method for drug-target interaction predictions and Alzheimer's disease drug repositions. *Applied Intelligence*, 52(1), 846–857. <https://doi.org/10.1007/s10489-021-02454-8>
- Wang, Steven, Wu, R., Lu, J., Jiang, Y., Huang, T., & Cai, Y.-D. (2022). Protein-protein interaction networks as miners of biological discovery. *Proteomics*, 22(15–16), e2100190. <https://doi.org/10.1002/pmic.202100190>
- Wang, X., Hripcsak, G., Markatou, M., & Friedman, C. (2009). Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. *Journal of the American Medical Informatics Association*, 16(3), 328–337. <https://doi.org/10.1197/jamia.M3028>
- Weissenrieder, J. S., Neighbors, J. D., Mailman, R. B., & Hohl, R. J. (2019). Cancer and the dopamine D2 receptor: A pharmacological perspective. *The Journal of Pharmacology and Experimental Therapeutics*, 370(1), 111–126. <https://doi.org/10.1124/jpet.119.256818>
- Wei, W.-Q., Cronin, R. M., Xu, H., Lasko, T. A., Bastarache, L., & Denny, J. C. (2013). Development and evaluation of an ensemble resource linking medications to their indications. *Journal of the American Medical Informatics Association*, 20(5), 954–961. <https://doi.org/10.1136/amiajnl-2012-001431>
- Wey, T., Blumstein, D. T., Shen, W., & Jordán, F. (2008). Social network analysis of animal behaviour: a promising tool for the study of sociality. *Animal Behaviour*, 75(2), 333–344. <https://doi.org/10.1016/j.anbehav.2007.06.020>
- Whirl-Carrillo, M., Huddart, R., Gong, L., Sangkuhl, K., Thorn, C. F., Whaley, R., & Klein, T. E. (2021). An Evidence-Based Framework for Evaluating

- Pharmacogenomics Knowledge for Personalized Medicine. *Clinical Pharmacology and Therapeutics*, 110(3), 563–572. <https://doi.org/10.1002/cpt.2350>
- Wieder, R., & Adam, N. (2022). Drug repositioning for cancer in the era of AI, big omics, and real-world data. *Critical Reviews in Oncology/Hematology*, 175, 103730. <https://doi.org/10.1016/j.critrevonc.2022.103730>
- Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., Assempour, N., Iynkkaran, I., Liu, Y., Maciejewski, A., Gale, N., Wilson, A., Chin, L., Cummings, R., Le, D., ... Wilson, M. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research*, 46(D1), D1074–D1082. <https://doi.org/10.1093/nar/gkx1037>
- Wu, C., Macleod, I., & Su, A. I. (2013). BioGPS and MyGene.info: organizing online, gene-centric information. *Nucleic Acids Research*, 41(Database issue), D561-5. <https://doi.org/10.1093/nar/gks1114>
- Wu, C.-H., Bai, L.-Y., Tsai, M.-H., Chu, P.-C., Chiu, C.-F., Chen, M. Y., Chiu, S.-J., Chiang, J.-H., & Weng, J.-R. (2016). Pharmacological exploitation of the phenothiazine antipsychotics to develop novel antitumor agents—A drug repurposing strategy. *Scientific Reports*, 6, 27540. <https://doi.org/10.1038/srep27540>
- Wu, M., Yi, H., & Ma, S. (2021). Vertical integration methods for gene expression data analysis. *Briefings in Bioinformatics*, 22(3). <https://doi.org/10.1093/bib/bbaa169>
- Wu, Q., Bagdad, Y., Taboureau, O., & Audouze, K. (2021). Capturing a comprehensive picture of biological events from adverse outcome pathways in the drug exposome. *Frontiers in Public Health*, 9, 763962. <https://doi.org/10.3389/fpubh.2021.763962>
- Wu, Y., Warner, J. L., Wang, L., Jiang, M., Xu, J., Chen, Q., Nian, H., Dai, Q., Du, X., Yang, P., Denny, J. C., Liu, H., & Xu, H. (2019). Discovery of noncancer drug

- effects on survival in electronic health records of patients with cancer: A new paradigm for drug repurposing. *JCO Clinical Cancer Informatics*, 3, 1–9.  
<https://doi.org/10.1200/CCI.19.00001>
- Xiong, Zhankun, Huang, F., Wang, Z., Liu, S., & Zhang, W. (2022). A multimodal framework for improving in silico drug repositioning with the prior knowledge from knowledge graphs. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(5), 2623–2631. <https://doi.org/10.1109/TCBB.2021.3103595>
- Xiong, Zhuang, Li, M., Yang, F., Ma, Y., Sang, J., Li, R., Li, Z., Zhang, Z., & Bao, Y. (2020). EWAS Data Hub: a resource of DNA methylation array data and metadata. *Nucleic Acids Research*, 48(D1), D890–D895. <https://doi.org/10.1093/nar/gkz840>
- Xu, C., Ai, D., Shi, D., Suo, S., Chen, X., Yan, Y., Cao, Y., Zhang, R., Sun, N., Chen, W., McDermott, J., Zhang, S., Zeng, Y., & Han, J.-D. J. (2018). Accurate Drug Repositioning through Non-tissue-Specific Core Signatures from Cancer Transcriptomes. *Cell Reports*, 25(2), 523-535.e5.  
<https://doi.org/10.1016/j.celrep.2018.09.031>
- Xu, L., Wang, Y.-Y., Huang, J., Chen, C.-Y., Wang, Z.-X., & Xie, H. (2020). Silver nanoparticles: Synthesis, medical applications and biosafety. *Theranostics*, 10(20), 8996–9031. <https://doi.org/10.7150/thno.45413>
- Xu, X., Yue, L., Li, B., Liu, Y., Wang, Y., Zhang, W., & Wang, L. (2022). DSGAT: predicting frequencies of drug side effects by graph attention networks. *Briefings in Bioinformatics*, 23(2). <https://doi.org/10.1093/bib/bbab586>
- Xu, Y., & McCord, R. P. (2022). Diagonal integration of multimodal single-cell data: potential pitfalls and paths forward. *Nature Communications*, 13(1), 3505.  
<https://doi.org/10.1038/s41467-022-31104-x>

- Yang, X., Coulombe-Huntington, J., Kang, S., Sheynkman, G. M., Hao, T., Richardson, A., Sun, S., Yang, F., Shen, Y. A., Murray, R. R., Spirohn, K., Begg, B. E., Duran-Frigola, M., MacWilliams, A., Pevzner, S. J., Zhong, Q., Wanamaker, S. A., Tam, S., Ghamsari, L., ... Vidal, M. (2016). Widespread expansion of protein interaction capabilities by alternative splicing. *Cell*, *164*(4), 805–817.  
<https://doi.org/10.1016/j.cell.2016.01.029>
- Yang, Y., Han, L., Yuan, Y., Li, J., Hei, N., & Liang, H. (2014). Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nature Communications*, *5*, 3231. <https://doi.org/10.1038/ncomms4231>
- Yan, J., Risacher, S. L., Shen, L., & Saykin, A. J. (2018). Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data. *Briefings in Bioinformatics*, *19*(6), 1370–1381.  
<https://doi.org/10.1093/bib/bbx066>
- Yoo, M., Shin, J., Kim, J., Ryall, K. A., Lee, K., Lee, S., Jeon, M., Kang, J., & Tan, A. C. (2015). DSigDB: drug signatures database for gene set analysis. *Bioinformatics*, *31*(18), 3069–3071. <https://doi.org/10.1093/bioinformatics/btv313>
- Yuan, L., Chen, L., Qian, K., Qian, G., Wu, C.-L., Wang, X., & Xiao, Y. (2017). Co-expression network analysis identified six hub genes in association with progression and prognosis in human clear cell renal cell carcinoma (ccRCC). *Genomics Data*, *14*, 132–140. <https://doi.org/10.1016/j.gdata.2017.10.006>
- Zame, W. R., Bica, I., Shen, C., Curth, A., Lee, H.-S., Bailey, S., Weatherall, J., Wright, D., Bretz, F., & van der Schaar, M. (2020). Machine learning for clinical trials in the era of COVID-19. *Statistics in Biopharmaceutical Research*, *12*(4), 506–517.  
<https://doi.org/10.1080/19466315.2020.1797867>



- Zeng, X., Zhu, S., Liu, X., Zhou, Y., Nussinov, R., & Cheng, F. (2019). deepDR: a network-based deep learning approach to in silico drug repositioning. *Bioinformatics*, *35*(24), 5191–5198. <https://doi.org/10.1093/bioinformatics/btz418>
- Zhang, F., Sun, B., Diao, X., Zhao, W., & Shu, T. (2021). Prediction of adverse drug reactions based on knowledge graph embedding. *BMC Medical Informatics and Decision Making*, *21*(1), 38. <https://doi.org/10.1186/s12911-021-01402-3>
- Zhang, R., Hristovski, D., Schutte, D., Kastrin, A., Fiszman, M., & Kilicoglu, H. (2021). Drug repurposing for COVID-19 via knowledge graph completion. *Journal of Biomedical Informatics*, *115*, 103696. <https://doi.org/10.1016/j.jbi.2021.103696>
- Zhang, Z., & Sejdíć, E. (2019). Radiological images and machine learning: Trends, perspectives, and prospects. *Computers in Biology and Medicine*, *108*, 354–370. <https://doi.org/10.1016/j.compbiomed.2019.02.017>
- Zhao, M., He, W., Tang, J., Zou, Q., & Guo, F. (2022). A hybrid deep learning framework for gene regulatory network inference from single-cell transcriptomic data. *Briefings in Bioinformatics*, *23*(2). <https://doi.org/10.1093/bib/bbab568>
- Zheng, X., Wang, B., Zhao, Y., Mao, S., & Tang, Y. (2021). A knowledge graph method for hazardous chemical management: Ontology design and entity identification. *Neurocomputing*, *430*, 104–111. <https://doi.org/10.1016/j.neucom.2020.10.095>
- Zhou, H., Cao, H., Matyunina, L., Shelby, M., Cassels, L., McDonald, J. F., & Skolnick, J. (2020). MEDICASCY: A Machine Learning Approach for Predicting Small-Molecule Drug Side Effects, Indications, Efficacy, and Modes of Action. *Molecular Pharmaceutics*, *17*(5), 1558–1574. <https://doi.org/10.1021/acs.molpharmaceut.9b01248>

- Zhou, M., Wang, Q., Zheng, C., John Rush, A., Volkow, N. D., & Xu, R. (2021). Drug repurposing for opioid use disorders: integration of computational prediction, clinical corroboration, and mechanism of action analyses. *Molecular Psychiatry*, *26*(9), 5286–5296. <https://doi.org/10.1038/s41380-020-01011-y>
- Zhou, Xu, Dai, E., Song, Q., Ma, X., Meng, Q., Jiang, Y., & Jiang, W. (2020). In silico drug repositioning based on drug-miRNA associations. *Briefings in Bioinformatics*, *21*(2), 498–510. <https://doi.org/10.1093/bib/bbz012>
- Zhou, XueZhong, Menche, J., Barabási, A.-L., & Sharma, A. (2014). Human symptoms-disease network. *Nature Communications*, *5*, 4212. <https://doi.org/10.1038/ncomms5212>
- Zhou, Y., Hou, Y., Shen, J., Huang, Y., Martin, W., & Cheng, F. (2020). Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. *Cell Discovery*, *6*, 14. <https://doi.org/10.1038/s41421-020-0153-3>
- Zhu, H., Zhang, J., Kim, M. T., Boison, A., Sedykh, A., & Moran, K. (2014). Big data in chemical toxicity research: the use of high-throughput screening assays to identify potential toxicants. *Chemical Research in Toxicology*, *27*(10), 1643–1651. <https://doi.org/10.1021/tx500145h>
- Zhu, Y., Che, C., Jin, B., Zhang, N., Su, C., & Wang, F. (2020). Knowledge-driven drug repurposing using a comprehensive drug knowledge graph. *Health Informatics Journal*, *26*(4), 2737–2750. <https://doi.org/10.1177/1460458220937101>
- Zielinski, J. M., Luke, J. J., Guglietta, S., & Krieg, C. (2021). High Throughput Multi-Omics Approaches for Clinical Trial Evaluation and Drug Discovery. *Frontiers in Immunology*, *12*, 590742. <https://doi.org/10.3389/fimmu.2021.590742>
- Zong, N., Wen, A., Moon, S., Fu, S., Wang, L., Zhao, Y., Yu, Y., Huang, M., Wang, Y., Zheng, G., Mielke, M. M., Cerhan, J. R., & Liu, H. (2022). Computational drug

repurposing based on electronic health records: a scoping review. *Npj Digital Medicine*, 5(1), 77. <https://doi.org/10.1038/s41746-022-00617-6>

# APPENDIX A – THE UNIFIED KNOWLEDGE SPACE

## The UKS – Materials and Methods

### Data Sources

**Table A 1** Data types and sources integrated into the UKS at this point of time. The data is grouped into 5 different categories: a) interaction: describes direct interactions between two nodes, such as protein protein interactions or drug interactions, b) regulation: describes a regulation relationship between two nodes, such as transcription factor gene regulation, c) functional: describes information about a nodes function or its mechanism of action, such as pathways or a compounds effect on the system, d) associations: describe relationships between nodes that are not of the previous three types, such as ontologies, and e) informative: contains additional knowledge about a node, such as different names, identifiers or sub-structures. To date the UKS contains information collected from more than 80 different databases and data collections. The last column describes if the data in the sources mainly contributes to the relationship layer, entity layer or both, as described in *Figure 8*.

Data Type	Data Sources	Data Category	Description	Contributes to
Gene product relationships	HIPPIE (Alanis-Lobato et al., 2017) HitPredict (López et al., 2015; Patil et al., 2011) HuRI (Luck et al., 2020) HI-union (Luck et al., 2020) Lit-BM (Luck et al., 2020) Yang-16 (X. Yang et al., 2016) HI-II-14 (Rolland et al., 2014) PINA (Du et al., 2021) MINT (Chatr-aryamontri et al., 2007)	Interaction regulation	Protein protein interactions, gene regulation (transcription factor, mirRNA)	Relationship Layer

	<p>PhosphoNetworks (Jianfei Hu et al., 2014)</p> <p>InnateDB (Breuer et al., 2013)</p> <p>SignalLink (Csabai et al., 2022; Fazekas et al., 2013)</p> <p>KEGG (M Kanehisa &amp; Goto, 2000; Kanehisa et al., 2017)</p> <p>Reactome (Jassal et al., 2020)</p> <p>TRRUST (Han et al., 2018)</p> <p>TargetScan (Agarwal et al., 2015)</p> <p>JASPAR (Castro-Mondragon et al., 2022; Sandelin et al., 2004)</p> <p>STRING (Szkarczyk et al., 2015, 2019)</p> <p>MiRTarBase (Huang et al., 2022)</p> <p>TransmiR (Z. Tong et al., 2019; J. Wang et al., 2010)</p>			
Associations – Compounds	<p>CTD (Davis et al., 2023)</p> <p>KEGG (M Kanehisa &amp; Goto, 2000; Kanehisa et al., 2017)</p> <p>OpenTargets (Ochoa et al., 2023)</p> <p>DrugBank (Wishart et al., 2018)</p>	Functional Associations interaction	Relationships between Chemicals and any node type, such as genes, phenotypes or other compounds. A compound can be a chemical, drug or engineered nanomaterial.	Relationship Layer

	<p>SMPDB (Frolkis et al., 2010; Jewison et al., 2014)</p> <p>STITCH (Szklarczyk et al., 2016)</p> <p>SIDER (Kuhn et al., 2016)</p> <p>MEDI (Wei et al., 2013)</p> <p>LabeledIn (Khare et al., 2014)</p> <p>BioSNAP<sup>11</sup> (Wishart et al., 2018)</p> <p>Zhou et al. (XueZhong Zhou et al., 2014)</p> <p>Wang et al. (F. Wang et al., 2014)</p> <p>Offsides<sup>12</sup></p> <p>Pharos (Kelleher et al., 2023)</p> <p>GuidetoPharmacology (Harding et al., 2022)</p> <p>DrugCombDB (H. Liu et al., 2020)</p> <p>RxNorm (S. J. Nelson et al., 2011)</p> <p>DSigDB (Yoo et al., 2015)</p>			
Associations - Phenotypes	<p>CTD (Davis et al., 2021)</p> <p>KEGG (M Kanehisa &amp; Goto, 2000; Kanehisa et al., 2017)</p> <p>SIDER (Kuhn et al., 2016)</p>	Functional associations	Relationships between phenotypes and any node type	Relationship Layer

<sup>11</sup> <https://snap.stanford.edu/biodata/datasets/10002/10002-ChG-Miner.html>

<sup>12</sup> <https://nsides.io/#offsides-and-twosides>

	<p>MEDI (Wei et al., 2013)</p> <p>LabeledIn (Khare et al., 2014)</p> <p>BioSNAP (Wishart et al., 2018)</p> <p>Zhou et al. (XueZhong Zhou et al., 2014)</p> <p>DisGeNet (Piñero et al., 2020)</p> <p>PsyGenNet (Gutiérrez-Sacristán et al., 2015, 2017)</p> <p>Wang et al. (F. Wang et al., 2014)</p> <p>Offsides<sup>13</sup></p> <p>Pharos (Kelleher et al., 2023)</p> <p>HPO (Köhler et al., 2021)</p> <p>NCBI ClinVar (Landrum et al., 2018)</p> <p>NCBI PheGenI (Sayers et al., 2022)</p> <p>PheWeb (Gagliano Taliun et al., 2020)</p> <p>Orphanet<sup>14</sup></p> <p>OMIM (Amberger et al., 2015, 2019)</p> <p>GWAS Catalog (Sollis et al., 2023)</p> <p>EWAS Data Hub (Zhuang Xiong et al., 2020)</p> <p>EWAS Catalog (Battram et al., 2022)</p>			
--	---	--	--	--

<sup>13</sup> <https://tatonetttilab.org/offsides/>

<sup>14</sup> [www.orpha.net/consor/cgi-bin/index.php](http://www.orpha.net/consor/cgi-bin/index.php)

	EWAS atlas (M. Li et al., 2019)			
Gene Sets	GO (The Gene Ontology Consortium, 2021) Reactome (Jassal et al., 2020) KEGG (M Kanehisa & Goto, 2000; Kanehisa et al., 2017) Wikipathway (Martens et al., 2021) PANTHER (Mi et al., 2021) MSigDB (Liberzon et al., 2011, 2015; Subramanian et al., 2005) SMPDB (Frolkis et al., 2010; Jewison et al., 2014)	functional	Collection of genes, associated to specific functions or phenotypes	Relationship Layer Entity Layer
Functional Cascades	KEGG (M Kanehisa & Goto, 2000; Kanehisa et al., 2017) Reactome (Jassal et al., 2020) Wikipathway (Martens et al., 2021) AOPwiki <sup>15</sup>	functional	MOA and how they are causative of each other.	Relationship Layer
Biological System	EWAS Data Hub (Zhuang Xiong et al., 2020) EWAS Catalog (Battram et al., 2022) EWAS atlas (M. Li et al., 2019)	informative	Information about organisms, tissues, cell types and cell lines.	Entity Layer

---

<sup>15</sup> aopwiki.org



	<p>Expression Atlas (Papatheodorou et al., 2018)</p> <p>ENCODE (Luo et al., 2020)</p> <p>Tabula Sapiens (Tabula Sapiens Consortium* et al., 2022)</p> <p>Cancer Cell Line Encyclopedia (Barretina et al., 2012)</p> <p>GTEx (GTEx Consortium, 2013)</p>			
Gene product information	<p>Ensembl (Cunningham et al., 2022)</p> <p>NCBI HomoloGen (Sayers et al., 2022)</p> <p>NCBI Entrez (Maglott et al., 2011; Sayers et al., 2022)</p> <p>GWAS Catalog (Sollis et al., 2023)</p> <p>EWAS Data Hub (Zhuang Xiong et al., 2020)</p> <p>EWAS Catalog (Battram et al., 2022)</p> <p>EWAS atlas (M. Li et al., 2019)</p> <p>InterPro (Paysan-Lafosse et al., 2023)</p>	informative	Additional information about a node, such as different identifiers, protein families, homologs	Entity Layer
Phenotype information	<p>NCBI MedGen (Sayers et al., 2022)</p> <p>ICD10<sup>16</sup></p>	informative	Additional information about phenotypes, such as different identifiers	Entity Layer

<sup>16</sup> [www.cdc.gov/nchs/icd/icd-10-cm.htm](http://www.cdc.gov/nchs/icd/icd-10-cm.htm)

	<p>ICD9<sup>17</sup>  HPO (Köhler et al., 2021)  Experimental Factor Ontology (Malone et al., 2010)  MONDO Disease Ontology (Mungall et al., 2017)  Disease Ontology (Schröml et al., 2023)  Ontology for Biomedical Investigations (Bandrowski et al., 2016)  Ontology for General Medical Science (Scheuermann et al., 2009)  Orphanet<sup>18</sup>  National Cancer Institute Thesaurus<sup>19</sup>  Mammalian Phenotype Ontology (Smith &amp; Eppig, 2009)  OpenTargets (Ochoa et al., 2023)</p>		<p>or relationships between phenotypes.</p>	
Compound information	<p>Food and Drug Administration product database<sup>20</sup>  NCBI PubChem (Kim et al., 2023)  ZINC20 (Irwin et al., 2020)</p>	informative	<p>Additional information about a compound, such as different identifiers or structural information</p>	Entity Layer

<sup>17</sup> [www.cdc.gov/nchs/icd/icd9.htm](http://www.cdc.gov/nchs/icd/icd9.htm)

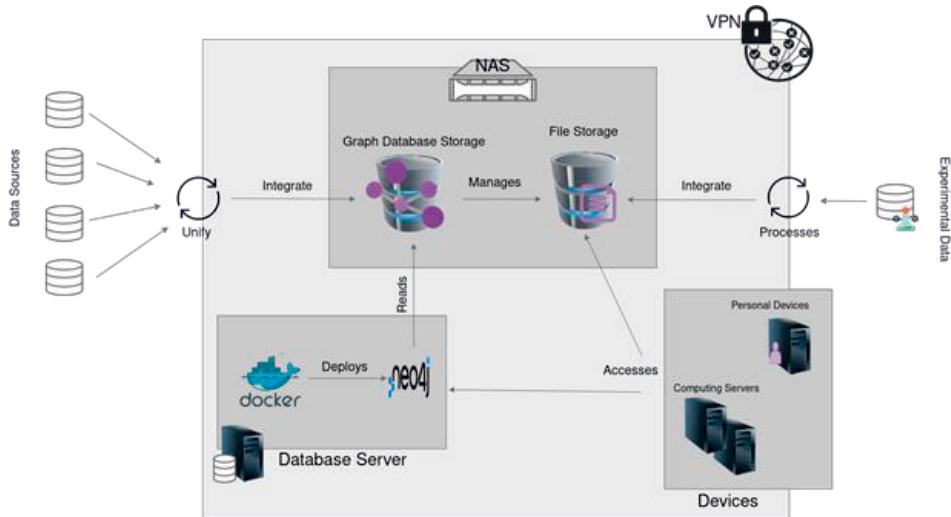
<sup>18</sup> [www.orpha.net](http://www.orpha.net)

<sup>19</sup> [ncithesaurus.nci.nih.gov/ncitbrowser/](http://ncithesaurus.nci.nih.gov/ncitbrowser/)

<sup>20</sup> <https://www.fda.gov/drugs>

	DrugTax (Preto et al., 2022)			
Experimental	<p>Open TG-GATEs (Igarashi et al., 2015)</p> <p>EWAS Data Hub (Zhuang Xiong et al., 2020)</p> <p>EWAS Catalog (Battram et al., 2022)</p> <p>EWAS atlas (M. Li et al., 2019)</p> <p>Expression Atlas (Papatheodorou et al., 2018)</p> <p>ENCODE (Luo et al., 2020)</p> <p>Tabula Sapiens (Tabula Sapiens Consortium* et al., 2022)</p> <p>Cancer Cell Line Encyclopedia (Barretina et al., 2012)</p> <p>GTEX (GTEX Consortium, 2013)</p> <p>Engineered Nanomaterial Transcriptomics Collection (Saarimäki et al., 2021)</p> <p>Human Protein Atlas (Thul &amp; Lindskog, 2018)</p>	functional	Such as gene expression, differential expression analysis (limma (Ritchie et al., 2015)). For most the data is stored in a file storage and its metadata is managed in the UKS.	Relationship Layer

## System Architecture



**Figure A 1** Data and System infrastructure used to manage and deploy the data. External data sources are processed and integrated into a graph database storage, hosted on a NAS (Network Attached Storage). Experimental data, which when processed, is not suitable to be hosted in a graph data model is stored in a file storage and its metadata is stored in the graph database storage. The file storage is hosted on a NAS, accessible from both computing servers and personal devices, when connected to the VPN (Virtual Private Network). The graph database storage is read and hosted with Neo4j, which is deployed via Docker on the database server. The deployed database management system can be accessed from both computing servers and personal devices when situated within the VPN. Access security is managed by the VPN access, which is due to the completely internal use and limited number of users enough. Also, while some of the data may be licensed, none of the hosted data are of sensitive nature.

## Indexing to Improve Query Performance

**Table A 2**

Custom created indices in the UKS. Indices are only created for nodes, since the default LOOKUP index covers node label and edge type searches and most of the expected UKS queries are node and relationship type focused, meaning edge attribute-based searches are not expected. Due to the more expensive (with respect of storage) nature of TEXT indices, they are only created for data points where substring searchers are expected, such as PHENOTYPE or CHEMICAL names. For ID based attributes equality searchers are expected, hence RANGE indices have been selected. To save storage, indices have only been created for highly queried or node types for which large amounts of data are available. The number of nodes per label are displayed in Table A 3.

Node Type	Index Attribute	Index Type	Expected searchers
GENE	Ensembl_ID	RANGE	=
GENE	Gene_symbol	TEXT	=, CONTAINS, STARTS WITH, ENDS WITH
VARIANT	CpG_site_ID SNP_ID	RANGE	=
GENE_EXPRESSION	Ensembl_ID	RANGE	=
CELL_TYPE	name	TEXT	=, CONTAINS, STARTS WITH, ENDS WITH
TISSUE	name	TEXT	=, CONTAINS, STARTS WITH, ENDS WITH
PHENOTYPE	HP_ID Concept_ID	RANGE	=
PHENOTYPE	name	TEXT	=, CONTAINS, STARTS WITH, ENDS WITH
AOP KEY_EVENT	name	TEXT	=, CONTAINS, STARTS WITH, ENDS WITH
CHEMICAL	name	TEXT	=, CONTAINS, STARTS WITH, ENDS WITH
PATHWAY	X_ID	RANGE	=
GO	GO_ID	RANGE	=
CHEMICAL	PubChem_CID PubChem_SID Zinc20_ID	RANGE	=
DATA_SET	name	TEXT	=, CONTAINS, STARTS WITH, ENDS WITH

## Main UKS Entities and their Identification System

**Table A 3** The main UKS node types and their selected identification system to which other identification systems are mapped.

Node Type	Main Identifier	Description
GENE	Ensembl ID	Established gene identification system, widespread use in the European research community. Good mapping APIs towards other systems exist, to integrate into the UKS Python's mygene API (C. Wu et al., 2013) is used.
CHEMICAL	PubChem CID & SID	Established identification system and provides a search API (Kim et al., 2023).
CHEMICAL:PATTERN	Canonical SMILES	Widespread representation of chemical structures, though if non canonical SMILES are used, they may differ for the same compound structure.
COMPOUND_CLASSIFICATION	name	As defined by source system.
PHENOTYPE	Concept ID	NCBI MedGen Concept ID (Sayers et al., 2022), provides an API and maps between multiple different identification systems.
ORGANISM	Taxonomy ID	Widespread use and easily accessible (Schoch et al., 2020).
PATHWAY	Internal ID of the different databases (KEGG (Kanehisa et al., 2017), Reactome (Jassal et al., 2020), Wikipathway (Martens et al., 2021) and SMPDB (Jewison et al., 2014))	Reactome, KEGG and Wikipathway are widely in use. SMPDB is the pathway library connected to DrugBank (Wishart et al., 2018) compounds. There is no direct mapping between pathways of different databases.
GO	Gene Ontology (Boyle et al., 2004) identifier	Widely in use.
PROTEIN_FAMILY	Panther ID (Mi et al., 2021) InterPro ID (Paysan-Lafosse et al., 2023)	Identification of the source system, due to independent definitions and categories of protein families.
GENE_SET	MSigDB ID (Liberzon et al., 2015)	Collection of gene sets as defined by source system.
VARIANT	SNP ID CpG site ID	Defined system, no mapping between the two identifiers.
ADVERSE_OUTCOME_PATHWAY KEY_EVENT	AOP ID	Identifier of source system. As defined by source system.

TISSUE	name	Custom curated, since no official widespread terminology/ ontology exists.
CELL_LINE CELL_TYPE	name	Custom curated, since no official widespread terminology/ ontology exists.
SAMPLE_DESCRIPTOR ETHNICITY AGE SEX	name	Custom curated, since no official widespread terminology/ ontology exists.
CELL_EXPERIMENTAL_DESCRIPTOR GROWTH LIFE_SPAN BASE_MEDIUM	name	Custom curated, since no official widespread terminology/ ontology exists.

## The UKS – Results

### The UKS as a Robust – Multidimensional Data Source

**Table A 4** Node labels and to date data point count in the UKS. To date there are ~68 million data points stored on the nodes in the UKS.

Main Node Label	Sub-labels	Number of Nodes to Date in UKS	Most common node attributes	Estimated Number of Data Points
ADVERSE_OUTCOME_PATHWAY (AOP)		412	Created Name AOP_ID	1,236
CELL_EXPERIMENTAL_DESCRIPTOR (CED)	BASE_MEDIUM GROWTH SAMPLING_SITE LIFE_SPAN	109	Created Name	218
CELL_LINE/ CELL_TYPE	DISEASED NORMAL CELL_LINE_TYPE	2,893	Created Name	5,786
CHEMICAL	DRUG DRUG_COMBINATION NANOMATERIAL (ENM) PATTERN RING STRESSOR MATERIAL	4,128,579	Created Name PubChem_CID SMILES	16,514,316
COMPOUND_CLASSIFICATION (CCL)	CLASS KINGDOM SUBCLASS	859	Created Name Classification_system	2,577

	SUPERCLASS			
DATA_SET	PROJECT STUDY	340	Created Name Description Provided_by Data_set_ID File_location	2,040
DATA_SET_T YPE		6	Created Name	12
GENE	MICRO_RNA PROTEIN_CODING PSEUDOGENES SNO_RNA LNC_RNA TRANSCRIPTION_FAC TOR	146,133	Created Ensembl_ID Gene_symbol Entrez_ID Biotype Ensembl_Protein_ID Ensembl_Transcript_I D Strand Start End Contig Assembly_name	1,753,596
GENE_EXPR SSION	HIGH MEDIUM LOW NOT_EXPRESSED	716,927	Created Ensembl_ID	1,433,854
GENE_SET		10,420	Created Name Description	31,260
GO	BIOLOGICAL_PROCES S CELLULAR_COMPO NENT MOLECULAR_FUNCTI ON	44,369	Created Name GO_ID	133,107
INFORMATIO N_ENTITY	CLINICAL_HISTORY MEASUREMENT	1 755	Created Name EFO_ID	5,265
KEY_EVENT	ADVERSE_OUTCOME MOLECULAR_INITIAT ING_EVENT	1 626	Created Name AOP_ID	4,878
ORGANISM		339	Created Name Taxonomy_ID	1,017
PATHWAY	DRUG_ACTION DRUG_METABOLISM METABOLIC PHYSIOLOGICAL	56,725	Created Name Source_ID	170,175



	PROTEIN SIGNALING			
PHENOTYPE	ABNORMALITY CLINICAL_COURSE DISEASE FINDING GROUP SIDE_EFFECT SYMPTOM	71,345	Created Name Concept_ID MESH_ID HPO_ID ICD10_ID	428,070
PROTEIN_FAMILY (PF)	ACTIVE_SITE BINDING_SITE CONSERVED_SITE HOMOLOGOUS_SUPERFAMILY HOMOLOGUE_GENE_GROUP PROTEIN_DOMAIN PTM REPEAT SUB_FAMILY	76,308	Created Name	152,616
PUBCHEM_FINGERPRINT (PCF)		881	Created Bit	1,762
SAMPLE_DESCRIPTOR (SD)	AGE_GROUP ETHNICITY SEX	35	Created Name	70
SPECIFIC_EXPERIMENT_CELL_LINE/ SPECIFIC_EXPERIMENT_CELL_TYPE (SEC)		4,318	Created Name Provided_by Experiment_ID	17,272
SPECIFIC_EXPERIMENT_TISSUE (SET)		54	Created Name Provided_by Experiment_ID	216
SPECIFIC_KEY_EVENT (SKE)		2,724	Created Name AOP_ID Event_AOP_ID AOP_AOP_ID	13,620
TISSUE	TISSUE_GROUP	169	Created Name	338
VARIANT		7,996,538	Created SNP_ID Allele Start End	47,979,228

			chromosome	
<b>Total</b>		<b>13,267,245</b>		<b>68,662,672</b>

**Table A 5** Edge types and to date data points stored on the relationships in the UKS. To date there are ~3.3 billion edge data points stored in the UKS.

Edge Type	Start Node Type	End Node Type	Number of Edges to date in the UKS	Most Common Edge Attributes	Estimated Number of Data Points
ACTIVATES	GENE CHEMICAL	GENE	100,015	Source Directed Created Downloaded	400,060
AFFECTS	CHEMICAL	GENE ORGANISM	1,175,742	Source Directed Created Downloaded	4,702,968
ASSOCIATED_WITH	CHEMICAL GO GENE PHENOTYP E CHEMICAL GENE PHENOTYP E GENE VARIANT CHEMICAL ORGANISM CHEMICAL CELL_LINE  PF PHENOTYP E	PHENOTYPE ORGANISM PATHWAY PATHWAY GO PHENOTYPE PHENOTYPE GO PHENOTYPE AOP AOP PATHWAY ORGANISM PHENOTYPE SD CED GO GO	50,670,061	Source Directed Created Downloaded	202,680,244
BINDS	CHEMICAL	GENE	4,780,759	Source Directed Created Downloaded	19,123,036
CATALYSIS	CHEMICAL	GENE	53,607	Source Directed Created Downloaded	214,428
CLASSIFIED_AS	PHENOTYP E	PHENOTYPE	2,074	Source Directed Created Downloaded	8,296
CONTAINS	DATA_SET	TISSUE CHEMICAL	47,122,432	Source Directed	188,489,728

	ENM CHEMICAL	DATA_SET CELL_LINE ORGANISM MATERIAL PATTERN		Created Downloaded	
DESCRIBES	CHEMICAL	PCF	556,040,688	Source Directed Created	1,668,122,064
DIFFERENTIAL_EXPRESSES	PHENOTYPE	GENE	1,331,236	Source Directed Created Downloaded FC_direction	6,656,180
DIRECTLY_LINKED_TO	SKE	SKE	2,192	Source Directed Created Downloaded	8,768
EXPRESSES	CHEMICAL CELL_TYPE CELL_LINE TISSUE	GENE	34,823,682	Source Directed Created Downloaded	139,294,728
EXPRESSION_CLASSIFICATION	CELL_TYPE CELL_LINE TISSUE	GENE	14,736,335	Source Directed Created Downloaded Probability_low Probability_medium Probability_high	88,418,010
HAS_AFFINITY	CHEMICAL	GENE	130,694	Source Directed Created Downloaded Affinity_score	653,470
HAS_CORE_MATERIAL	ENM	MATERIAL	559	Source Directed Created Downloaded	2,236
HAS_MATURE_SEQUENCE	MICRO_RNA	MICRO_RNA	21,032	Source Directed Created Downloaded	84,128
HAS_METASTASIS_IN	CELL_LINE	TISSUE	105	Source Directed Created	420

				Downloaded	
HAS_SAMPLING_SITE	CELL_LINE	SAMPLING_SITE	127	Source Directed Created Downloaded	508
HAS_SHAPE	ENM	GEOMETRY	551	Source Directed Created Downloaded	1,653
HAS_SIDE_EFFECT	CHEMICAL	PHENOTYPE	1,415,012	Source Directed Created Downloaded reporting_frequency	7,075,060
INDIRECTLY_LEADS_TO	SPECIFIC_KEY_EVENT	SPECIFIC_KEY_EVENT	299	Source Directed Created Downloaded	1,196
INHERITS_FROM	PHENOTYPE	PHENOTYPE	8,242	Source Directed Created Downloaded	32,968
INHIBITS	CHEMICAL	GENE	111,173	Source Directed Created Downloaded	444,692
INTERACTS_WITH	CHEMICAL	CHEMICAL_GENE	16,327,003	Source Directed Created Downloaded	65,308,012
IS_A	GO	GO	70,938	Source Directed Created Downloaded	283,752
IS_CAPABLE_OF	GO	GO	511	Source Directed Created Downloaded	2,044
IS_CAPABLE_OF_PART_OF	GO	GO	300	Source Directed Created Downloaded	1,200
IS_CAUSATIVE_OF	GENE CHEMICAL	PHENOTYPE	9,812	Source Directed Created Downloaded	39,248
IS_CHILD_OF	PF_GENE_SET	PF_GENE_SET	785,195	Source Directed	3,140,780

	SKE SET SEC CELL_LINE  GENE_EXP RESSION TISSUE ETHNICITY AGE_GROU P PHENOTYP E CELL_TYPE CCL	KEY_EVENT TISSUE CELL_LINE CELL_LINE_ TYPE GENE  TISSUE_GRO UPETHNICIT Y AGE_GROUP PHENOTYPE CELL_TYPE CCL		Created Downloaded	
IS_EVENT_OF	KEY_EVEN T	AOP	5,417	Source Directed Created Downloaded	21,668
IS_EXPOSED_ ON	CHEMICAL	TISSUE CELL_LINE ORGANISM	3,259	Source Directed Created Downloaded	21,668
IS_GENE_OF	ORGANISM	GENE	134,848	Source Directed Created	404,544
IS_ MANIFESTATI ON_OF	PHENOTYP E	PHENOTYPE	91,940	Source Directed Created Downloaded	367,760
IS_MAPPED_T O	KEY_EVEN T	PATHWAY GO PHENOTYPE GENE	4,075	Source Directed Created Downloaded	16,300
IS_ORTHOLO G_OF	GENE	GENE	66,413	Source Directed Created Downloaded Homology_t ype	332,065
IS_PARALOG_ OF	GENE	GENE	777,217	Source Directed Created Downloaded Homology_t ype	3,886,085
IS_PART_OF	GO CHEMICAL GENE	GO DATA_SET	15,655	Source Directed Created	62,620

IS_PATHWAY_OF	PATHWAY	ORGANISM	8,025	Source Directed Created Downloaded	24,075
IS_PRESENT	VARIANT	TISSUE CELL_TYPE	2,093,740	Source Directed Created Downloaded	8,374,960
IS_RELATED_TO	PHENOTYPE	PHENOTYPE	3,210	Source Directed Created Downloaded	12,840
IS_VARIANT_OF	VARIANT	GENE	6,746,801	Source Directed Created	20,240,403
MEDICATES	CHEMICAL	PHENOTYPE	27,877	Source Directed Created Downloaded	111,508
MEMBER_OF	GENE  CHEMICAL	PF GENE_SET DRUG_ COMBINATIO N CCL	23,932,237	Source Directed Created	719,796,711
NEGATIVELY_REGULATES	GO	GO	3,126	Source Directed Created Downloaded	12,504
NOT_EXPRESSES	CELL_TYPE	GENE	25,613,578	Source Directed Created	76,840,734
OCCURES_IN	GO	GO	300	Source Directed Created Downloaded	1,200
POSITIVELY_REGULATES	GO	GO	3,112	Source Directed Created Downloaded	12,448
P_P_INTERACTION	GENE	GENE	15,591,537	Source Directed Created Downloaded	62,366,148
REACTION	CHEMICAL	GENE	42,146	Source Directed Created Downloaded	168,584

REGULATES	GENE GO	GENE GO	858,966	Source Directed Created Downloaded regulation_ty pe	4,294,830
REPRESSES	GENE	GENE	711	Source Directed Created Downloaded	2,844
TARGETS	CHEMICAL	GENE	38,168	Source Directed Created Downloaded action	190,840
<b>TOTAL</b>			<b>1,021,782,734</b>		<b>3,292,744,584</b>



**Table A 6** To date data points stored in the file storage managed by the UKS. If a data set varies in their number of genes or samples across the collection an estimate is used.

Data Type	Data Set	Number of Sub Datasets	Measured Genes	Number Analysis/Samples	Estimated Data Points
Differential Expression Analysis with Limma	Open TG-GATEs (Igarashi et al., 2015)	1	19,939	938	18,702,782
Differential Expression Analysis with Limma	Engineered Nanomaterial Transcriptomics Collection (Saarimäki et al., 2021)	106	20,000	100	212,000,000
Gene Expression	Engineered Nanomaterial Transcriptomics Collection (Saarimäki et al., 2021)	106	20,000	100	212,000,000
Gene Expression	CCLE (Barretina et al., 2012)		53,827	1,405	75,626,935
Gene Expression	Tabula Sapiens (Tabula Sapiens Consortium* et al., 2022)		58,559	472	27,639,848
Gene Expression	Human Protein Atlas (Thul & Lindskog, 2018)		19,553	65	1,270,945
Gene Expression	ENCODE (Luo et al., 2020)		58,426	240	14,022,240
Gene Expression	Expression Atlas (Papatheodorou et al., 2018)	3	50,000	400	60,000,000
<b>Total</b>					<b>621,262,750</b>



# PUBLICATION

|

## **Integrated network analysis reveals new genes suggesting COVID-19 chronic effects and treatment**

Alisa Pavel, Giusy del Giudice, Antonio Federico, Antonio Di Lieto, Pia A S Kinaret, Angela Serra, Dario Greco

Briefings in Bioinformatics, Volume 22, Issue 2, March 2021, Pages 1430–1441  
<https://doi.org/10.1093/bib/bbaa417>

**Publication is licensed under a Creative Commons Attribution 4.0  
International License CC-BY-NC-ND**



# Integrated network analysis reveals new genes suggesting COVID-19 chronic effects and treatment

Alisa Pavel<sup>†</sup>, Giusy del Giudice<sup>†</sup>, Antonio Federico, Antonio Di Lieto, Pia A.S. Kinaret, Angela Serra and Dario Greco 

Corresponding author: Dario Greco. Faculty of Medicine and Health Technology, Arvo Ylpön katu 34, 33014 Tampere, Finland; BioMediTech Institute, Tampere University, Tampere, Finland; Institute of Biotechnology, University of Helsinki, Helsinki, Finland; The Finnish Centre for Alternative Methods (FICAM), Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland; Tel: 0503182106; E-mail: dario.greco@tuni.fi

<sup>†</sup>These authors contributed equally to this work.

## Abstract

The COVID-19 disease led to an unprecedented health emergency, still ongoing worldwide. Given the lack of a vaccine or a clear therapeutic strategy to counteract the infection as well as its secondary effects, there is currently a pressing need to generate new insights into the SARS-CoV-2 induced host response. Biomedical data can help to investigate new aspects of the COVID-19 pathogenesis, but source heterogeneity represents a major drawback and limitation. In this work, we applied data integration methods to develop a Unified Knowledge Space (UKS) and used it to identify a new set of genes associated with SARS-CoV-2 host response, both *in vitro* and *in vivo*. Functional analysis of these genes reveals possible long-term systemic effects of the infection, such as vascular remodelling and fibrosis. Finally, we identified a set of potentially relevant drugs targeting proteins involved in multiple steps of the host response to the virus.

**Key words:** COVID-19; SARS-CoV-2; coronavirus; virus–host interaction; unified knowledge space; data integration; multi-layer network analysis; drug targeting; drug repositioning

**Alisa Pavel** is a Ph.D. student in Prof. Dario Greco's group at Tampere University. She holds a M.Sc. in Computer Science (2019) from the University of Edinburgh. Her areas of expertise are biological network modelling and their analysis, as well as big data modeling for toxicogenomics and pharmacology. **Giusy del Giudice** is Ph.D. student in Prof. Dario Greco's group at the Tampere University. She holds a M.Sc. in Medical Biotechnology (2019). Her areas of expertise are the analysis and modelling of toxicogenomics and pharmacogenomics data, and systems immunology.

**Dr. Antonio Federico** works as a postdoctoral researcher in Prof. Dario Greco's group at the University of Tampere. Dr. Federico holds a master degree in Molecular Biology (2014) and an international PhD with a dissertation about multi-layer integration of molecular networks to uncover the mechanisms of drug sensitivity of cancer-related genes for the development of tailored pharmacological therapies (2019). His research interests lie in network analysis, predictive pharmacology, systems biology and multi-omics data analysis and integration.

**Antonio Di Lieto** is MD, Ph.D. specialist in psychiatry. He obtained his Ph.D. in Behavioural science and learning processes. He has previously researched the biological correlates of psychiatric syndromes and molecular pharmacology. He is currently working as senior consultant psychiatrist at the Department of Forensic Psychiatry of the University of Aarhus, Denmark.

**Dr. Pia A.S. Kinaret** is senior postdoctoral researcher in Prof. Dario Greco's group at the University of Helsinki. Kinaret has degrees in biotechnology (2009) and bioinformatics (2012), and she obtained the Ph.D. in genetics (2017). Her current research focuses on immunotoxic effects and immunomodulatory potential of manufactured nanomaterials. In her work she utilizes *in vivo*, *in vitro* and systems biology approaches. She currently receives funding from Orion Research Foundation sr.

**Dr. Angela Serra** is a senior postdoctoral researcher in Prof. Dario Greco's group at the University of Tampere. Serra has a master's degree in Computer Science (2013), and she obtained the Ph.D. in Computer Science and Information Engineering with a dissertation on multi-view learning and data integration approaches for multi-omics data (2017). Her current research focuses on data modelling, machine learning, network inference, data integration for toxicogenomics, bioinformatics and cheminformatics.

**Prof. Dario Greco** is professor of bioinformatics at the Faculty of Medicine and Health Technology, Tampere University and principal investigator at the Institute of Biotechnology, University of Helsinki, Finland. To date, he authored over 130 articles in peer reviewed journals in the areas of nanotoxicology, systems toxicology, toxicogenomics, predictive toxicology, data modelling and bioinformatics. He currently receives fundings from the Academy of Finland, Business Finland and the EU (H2020 and IM2 programs).

**Submitted:** 3 August 2020; **Received (in revised form):** 13 November 2020

© The Author(s) 2021. Published by Oxford University Press.

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

## Introduction

The newly identified coronavirus SARS-CoV-2 is responsible for a pandemic form of respiratory tract infection currently ongoing worldwide. Even if most patients remain asymptomatic or show mild symptoms, some develop complications, such as severe pneumonia and acute respiratory distress syndrome (ARDS) [1, 2]. Furthermore, systemic complications, such as cardiovascular disorders, persistent lung injuries and possibly fibrosis are rapidly emerging as key threats in addition to the respiratory syndrome. Restrictive measures have been adopted to slow down the spreading of the virus; however, it is expected that the infection will remain entrenched in the population for years [3].

To date, no approved vaccine is yet available and some therapeutic strategies have been proposed to control the clinical outcomes of the infection [4, 5]. Currently, a great effort is being made by the scientific community in order to develop new therapeutic approaches as well as to understand the molecular events characterizing the host response to SARS-CoV-2 infection. SARS-CoV-2 infects the cells via the angiotensin converting enzyme 2 (ACE2) receptor-mediated endocytosis [6]. ACE2 is expressed in several organs and cell types, such as lung, heart, kidney, intestine and endothelial cells, which further raises concerns about possible ectopic effects of the infection [7].

Molecular characterization of infected tissues and cells can elucidate key potential molecular targets involved in the pathogenesis of COVID-19. To this end, for instance, Gordon *et al.* [8] applied mass spectrometry to identify SARS-CoV-2 human protein interactors. These proteins can be considered as the first responders to the virus, acting upstream in the host response to the infection. Moreover, transcriptomic data of infected lungs and cell types are already publicly available [9]. On the contrary, the genes derived from the transcriptomic data represent late effectors in the host immune response. Nonetheless, a knowledge gap exists to link the first host responses to the virus with the subsequent phenotypic alterations. In this work, we hypothesize that genes linking the upstream interactors and downstream effectors are involved in the transduction and amplification of the host response to the virus and can therefore represent a new set of potentially relevant genes. Developing computational methods that are able to infer such missing information is of extreme importance, especially in situations where there are limited data available, such as in the COVID-19 disease. Moreover, a deeper understanding of the underlying molecular responses is required in order to develop suitable treatment methods and prepare for possible long-term effects. This gap could be filled by exploiting the large amount of biomedical data accumulated in recent years. However, the use of this information is currently hampered by the heterogeneity of data formats scattered across multiple repositories [10–12]. In this study, we applied scalable and flexible data integration methods to develop a robust compendium of molecular knowledge, the Unified Knowledge Space (UKS). Knowledge graphs (KGs) are large data structures that model different entities, their properties and relationships [13–16]. KGs allow to integrate multiple data from diverse domains and repositories into a common space. In this way, KGs facilitate the organization of information in a structured manner and allow to visualize and retrieve complex relationships between different entities derived from multiple sources. Another purpose of KGs is the generation of currently unknown facts, which can be inferred from existing links in the KG. In the domain of biology, KGs have for example been used in drug repositioning [17, 18] or to infer disease-biomolecule associations [19, 20]. In our UKS, nodes can be genes, gene

products or drugs, while edges represent different relationships between the entities. The UKS is created by combining homogeneous with heterogeneous network integration methods. Homogeneous network integration combines different networks with the same node (type) but different edges, merging them into a single network (e.g. combining multiple protein–protein interactions (PPI) networks), while heterogeneous network integration aims at connecting networks with different node (types) (e.g. gene–drug target networks with a gene–gene network) [21].

The expansion of the PPI network through other data types to construct a heterogeneous network has been previously applied in a variety of contexts [20, 22, 23]; Davis and Chawla [23] constructed a phenotypic-disease network merged with a genetic-disease network to investigate disease comorbidities, while Goh *et al.* [20] built a network linking genetic disorders with known disease genes to investigate the role of disease genes in the human interactome. A detailed review about different network data integration methods and their application is provided by Gligorijević and Pržulj [21]. While previous studies aimed at constructing a homogeneous or heterogeneous network for a specific case study, we built an expandable and flexible data structure. Consequently, high-quality networks can be inferred (homogeneous and/or heterogeneous networks can be retrieved). This allows the UKS to be used in a wide variety of different studies in the future.

We analysed the UKS and retrieved a novel set of genes potentially associated with the molecular host response to SARS-CoV-2 infection. The functional characterization of this new set of genes allows us to describe possible unpredicted long-term complications of the COVID-19 disease, as well as to suggest repositioning of some already approved drugs.

## Methods

The proposed methodology aims at giving insights into the possible mechanistic aspects of the SARS-CoV-2 infection and host response through the construction of a Unified Knowledge Space. Combining knowledge about viral physical interactor human proteins and transcriptomic studies into a single knowledge space allows to gain new valuable insights about the mechanisms underlying COVID-19. By further expanding the UKS with information about drug targets, valuable novel knowledge regarding multiple facets of the SARS-CoV-2 infection can be generated. We define the UKS as a knowledge graph constituted through multiple network layers [24, 25], where nodes are representing either gene (products) or drugs, and edges represent either direct known physical gene–gene interactions or drug–gene target relationships. The UKS comprises all human protein-coding genes as retrieved from Ensembl [26], known physical interactions of their associated proteins as well as all known drug target relationships. Our whole applied pipeline, including data retrieval, processing and knowledge extracted, is outlined as pseudocode in the Supplementary File S1 available online at <https://academic.oup.com/bib>.

## Data collection

### Viral interactors

Genes known to be physically interacting with the viral components of SARS-CoV-2 were retrieved from [8]. These genes are involved in the first events of the host response upon viral infection.

## Transcriptomics data

The gene expression data of human lung biopsies of SARS-CoV-2 infected patients and SARS-CoV-2 infected cell lines were retrieved from the Gene Expression Omnibus (GEO) repository (GEO ID GSE147507) [9]. The dataset only consisted of one time point, and RNA was extracted 24 h after the infection. In this work, we analyzed five different experimental conditions contained in the GEO dataset: human lung biopsies of SARS-CoV-2 infected patients and uninfected control; A549 cell line infected with SARS-CoV-2; A549 cell line infected with SARS-CoV-2 over-expressing ACE2; Calu-3 cells infected with SARS-CoV-2; NHBE cell line infected with SARS-CoV-2. For each of the cell lines, the mock treated lines were collected to be used as controls for the expression analysis.

## Transcriptomics data analysis (DE gene set identification)

Gene expression analysis was carried out starting from the raw counts provided within the GEO record. Low read counts were filtered by applying the proportion test method implemented within the NOISeq Bioconductor package [27]. Filtered counts were normalized through the upper quartile method implemented in the NOISeq package. Differential expression analysis was carried out by using the DESeq2 Bioconductor package [28], while p-values were adjusted through the Benjamini-Hochberg method [29]. The pre-processed expression matrices are reported in the Supplementary Files S2–S5 available online at <https://academic.oup.com/bib>.

## UKS construction and PPI network retrieval

Known human protein coding genes were retrieved from Ensembl (Assembly: GRCh38) [26], which represent the base of the developed UKS. Known protein–protein interactions were retrieved from HIPPIE (downloaded 28/10/2019) [30], HitPredict [31, 32] (downloaded 04/11/2019), KEGG (downloaded 08/12/2019) [33] and STRING (downloaded 23/02/2020) [34]. We combined these PPI networks into a unique homogeneous network, by mapping the proteins to their associated genes. The edges were weighted based on an interaction source support score, where an edge weight of 1 indicates source support by 100% of the collected sources. This is important, since it has been shown that there is a high variance between links in PPI networks, in terms of quality of the determined interactions (e.g. experimental based versus literature based). Therefore, the confidence of the interaction varies widely and, in addition, links between proteins may be missing [35–37]. To reduce the data quality bias, we consider source support for each edge as important to reveal a high confidence subnetwork from the homogeneously merged PPI network. This approach is similar to the robust PPI network construction approach suggested by Martha et al. [38]. Drug target information was collected from DrugBank (downloaded 22/04/2020) [39] and Open Targets (downloaded 15/02/2019) [40] and integrated into the UKS. The data contained in these two sources are merged into a single data layer by means of mapping drugs to PubChem CIDs or SIDs [41], again conserving source information. In order to link the data accurately to the previously discussed data layers, gene symbols are mapped to Ensembl Gene IDs [26] through [mygene.info](http://mygene.info) [42, 43]. To provide a highly flexible data provisioning system, the UKS is stored as a graph database in Neo4j 4.0 (<https://neo4j.com/>), which allows to edit, retrieve and add new data as needed. The

complete UKS contains 20 793 human protein coding genes, which are interlinked by 5 941 639 edges, representing physical known interactions. Additional 7099 drugs are linked through 22 973 edges to their genetic targets. To construct a high-quality gene–gene network, gene–gene relationships, associated with a source support score of at least 0.75, are queried from the UKS and used to construct a single layer gene–gene network, which is represented as a Python NetworkX graph [44]. The final gene–gene network is made up of 20 793 nodes, representing Ensembl gene IDs, interlinked by 132 244 high-quality edges, describing interactions between the gene's associated proteins.

## Identification of intermediate genes through shortest paths

In order to identify the relevant genes that may have a crucial role in the progression of SARS-CoV-2 infection, the shortest paths between the physical interacting (PI) and the differentially expressed (DE) gene sets were computed. Shortest path analysis is a method to link two sets of nodes of interest and identify interactor nodes between them. On a graph  $G = (V, E)$ , where  $V$  is the set of nodes and  $E$  is the set of edges, a shortest path between  $v_i$  and  $v_j$  is defined as the path between  $v_i$  and  $v_j$  requiring the least effort. In an unweighted network, this translates into the least number of steps to be taken to connect  $v_j$  and  $v_i$  [45]. The shortest path analysis was performed for each group of DE genes identified in each biological system.

The shortest paths were retrieved with Python NetworkX [44], `shortest_path()` function, by running Dijkstra's shortest path algorithm [45] between all possible pairs of PI and DE genes (Python 3.6.9, NetworkX 2.3). All edges were considered to have equal weight, meaning that only the number of steps was considered when running the algorithm. Only paths consisting of at least one intermediate gene (IN) (path length  $>1$ ) were considered during further analysis.

For each gene in the PPI network, its occurrences as an IN between PI and DE was estimated separately for each experimental class and statistically significant enriched IN genes were identified. Hypergeometric test was performed by comparing the IN frequencies identified in the shortest paths of interest with their occurrences on all possible shortest paths in the complete gene–gene network. By estimating statistical significance of each visited intermediate node, only intermediate nodes that are relevant in linking the previously defined sets of key nodes (DE and PI) are considered. Adjusted p-values were estimated by applying the Benjamin and Hochberg multiple testing correction [29]. The nominal p-values were calculated with Python's SciPy package [46] and the adjusted p-values were estimated based on Python's statsmodels package [47] (SciPy 1.3.2, statsmodels 0.11.1).

## Pathway enrichment analysis

In order to functionally characterize the lists of PIs, INs, and DEs, pathway enrichment analyses were performed using the Wikipathway 2019 Human database through the EnrichR online tool [48, 49]. The enriched pathways were visualized by means of the FunMapOne tool [50].

## Gene ranking

In order to evaluate the overall most common genes crossed in the shortest paths, for each in vivo and in vitro system, only statistically significant genes were selected and ranked

according to the intermediate gene count value. The five lists were given as an input to the Borda function of the TopKList R package [51], to calculate the Borda scores and rank the genes according to the median function.

### Identification of relevant drugs

In order to highlight drugs that could simultaneously affect multiple steps of the host response to SARS-CoV-2, we retrieved from the UKS the list of drugs targeting genes in the PI, IN and DE sets and retrieved the set contained in their intersection.

## Results and discussion

### A novel set of genes involved in the pathogenesis of COVID-19 can be retrieved from multi-scale molecular network analysis

The Unified knowledge space (UKS) defined in this work has been generated by integrating multiple data sets containing protein–protein interaction (PPI) information as well as drug–target relationships. By querying the UKS, we derived a network of 20 793 human protein coding genes, represented as nodes, and 132 244 edges, representing the physical interaction relationships existing between the proteins encoded by the UKS gene nodes. These interactions were integrated from four data sources and stored in the UKS together with a data support score, representing the number of sources in which the connections are present. In order to have a reliable structure of the network, we selected only edges supported in at least three out of four sources (see section ‘UKS Construction and PPI Network Retrieval’ for more details). The UKS network was further extended with gene–drug information by adding 7099 drug nodes that are linked to their target gene nodes through 22 973 edges (Figure 1A). We systematically mapped the SARS-CoV-2 physical interacting (PI) genes and the differentially expressed (DE) genes in multiple biological systems infected by SARS-CoV-2 [9] (Figure 1B).

A set of human proteins has been recently described by Gordon *et al.* as physical interactors of the SARS-CoV-2 viral components [8]. We considered these as the first set of proteins involved in the host response to a SARS-CoV-2 infection. On the other hand, we considered the differentially expressed genes retrieved from transcriptomic analysis of infected *in vivo* (infected versus healthy human lung biopsies) and *in vitro* (infected versus mock CALU-3, A549, A549 overexpressing ACE2, and NHBE cell lines) systems, as late effectors associated with the COVID-19 pathological phenotype. In order to identify the relationships between the first interactors of SARS-CoV-2 (PI gene set) and the late effectors (DE gene set), a third set of genes, located in the shortest path between each possible pair of (PI-DE genes) was retrieved.

The concept of shortest paths has already been widely applied in the analysis of biological networks and has yielded biologically relevant results [52–54]. Du *et al.* [52] mapped differentially expressed genes onto a PPI network and successfully identified transcription factors linking a cancer gene to its differentially expressed genes. Simões *et al.* [53] applied a similar strategy in order to identify genes associated to complex diseases.

In our study, we use the concept of shortest paths to investigate the set of genes linking the genes directly interacting with viral components and the ones whose transcription is altered by the induced host response. From a kinetics perspective, the

first set of genes (PI) can be assumed to have a role in the first molecular events upon viral exposure; on the contrary, modulation of the expression of the late effector genes (DE) is associated with cellular and, ultimately, systemic response to the infection. We, therefore, assumed that genes in the shortest paths can be involved in the transduction and amplification of the host response. In this light, the intermediate genes can better explain the chain of the molecular events characterizing the response to SARS-CoV-2, as well as can represent another important set of therapeutic targets.

For each *in vitro* and *in vivo* system analyzed, we named as intermediate genes (IN gene set), all the genes, not belonging to either the PI nor the DE gene sets, significantly overrepresented ( $P$ -value  $\leq 0.05$ ) in the shortest paths.

In contrast with the heterogeneity of the DE gene set sizes, the number of intermediate genes is comparable among the different biological systems (Figure 2). Overall, we observe a progressive increase in the size of the gene sets when going from the first interactors (PI), through the intermediate genes (IN), to the effector pathways genes (DE), suggesting the role of the intermediate genes in propagating the host response mechanisms to the virus entry. The human bronchial epithelial cells (NHBE), on the contrary, was the only dataset showing a decreasing trend from the PI to the DE gene set. This is probably due to the smaller number of differentially expressed genes, which can be associated with the lower permissiveness of the NHBE cell line.

### Functional characterization of intermediate gene set reveals possible long-term effects of COVID-19 disease

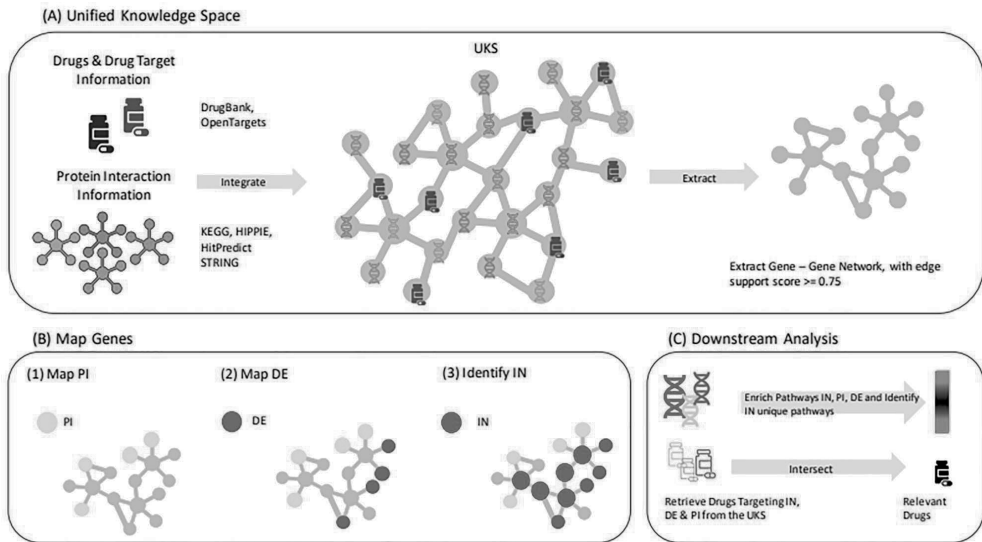
In order to characterize the IN gene set, we performed pathway enrichment analysis independently for each *in vivo* and *in vitro* biological system analyzed. Moreover, we compared the pathways over-represented in the IN set with the ones over-represented in the PI and DE genes, respectively, in order to identify specific biological functions, which could fill the gap between the early molecular interaction events and the downstream transcriptomic host response (Figure 3).

As expected, PI genes specifically enriched pathways related to viral infections, such as Ebola Virus pathway on Host and Dual hijack model of Vif in HIV infection. Not surprisingly, cilia associated pathways were also enriched, since epithelial cells are the first ones to encounter SARS-CoV-2 in the respiratory system. These pathways are also well represented in the IN genes, while they are not significantly enriched in the DE gene set.

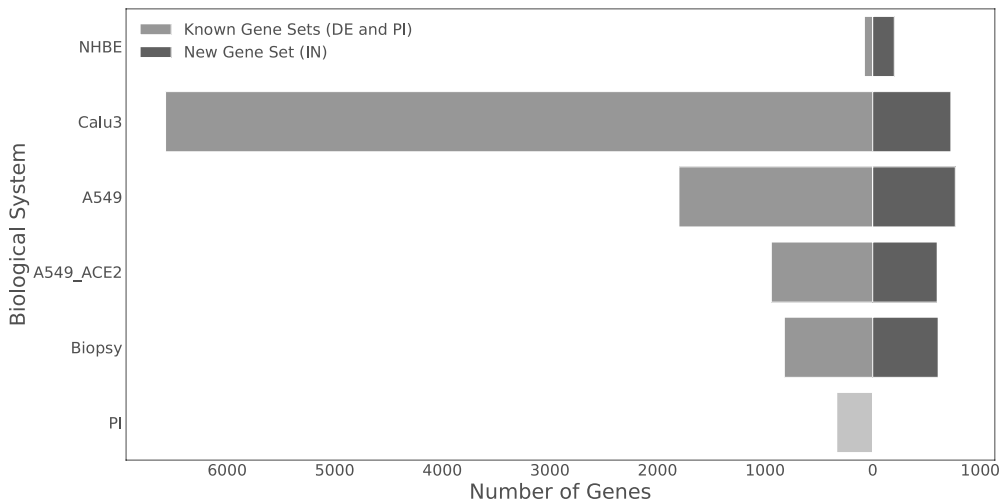
Metabolic pathways are present in all the gene sets (PI, IN and DE), with the oxidative phosphorylation and lipid metabolism being the most affected functions. Viral infections are known to induce a global metabolic alteration of the cell and, in particular, lipids play a pivotal role in facilitating viral replication [55].

DE genes specifically enriched immune system related pathways, which were not represented in the PI gene set and minorly represented in the IN set. Some of the main effector molecules involved in the cytokine storm observed in COVID-19 were present in the enriched immune pathways (e.g.  $INF\gamma$ , TNF, IL-1 $\beta$  and other chemokines) as well as the NF $\kappa$ B transcription factor pathway (Figure 3) [6, 56]. Interestingly, interferon response was retrieved as significantly over-represented both in IN and DE genes. However, type I interferon was specifically enriched in the IN set, whereas type II was enriched in the DE set only. Interferon gamma, the only type II interferon, is one of the genes involved in the cytokine storm [56]. On the contrary, type I

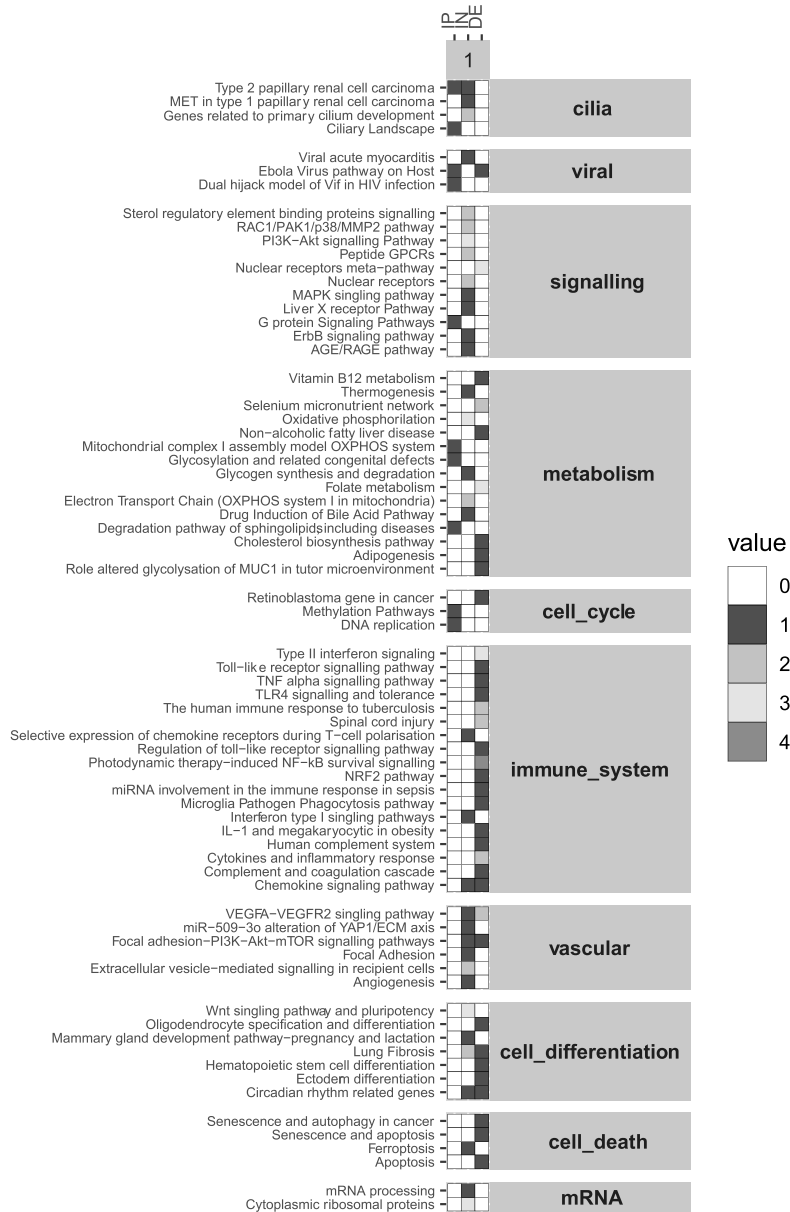




**Figure 1.** Scheme of the analytical framework. Data from multiple protein–protein interaction (PPI) sources (KEGG, HIPPIE, HitPredict and STRING) were collected and mapped to their corresponding Ensembl Gene IDs. The PPI network was further integrated with drug–target information, derived from DrugBank and OpenTargets, to form the UKS (A, left). A robust gene–gene network was extracted from the UKS, where only edges supported by at least three of the merged PPI networks were included (A, right). PI and DE genes were mapped onto the extracted gene–gene network and intermediate genes (IN) were identified by means of shortest paths between each possible pair of PI and DE (B). Pathway enrichment analysis was performed for all three gene sets. Drugs that have targets in all three gene sets (PI, DE and IN) were selected and classified as ‘relevant drugs’ (C).



**Figure 2.** Number of known gene sets (DE and PI) and new gene set (IN) in each biological system. The number of differentially expressed genes (DE), and the new retrieved set of intermediate genes (IN), are compared in each *in vitro* and *in vivo* system. The samples (biological system) derived from public transcriptomics dataset comprising a lung biopsy and four different cell lines infected with the virus: the transformed cell lines A549 (adenocarcinomic human alveolar basal epithelial cells) and CALU-3 (human lung cancer epithelial cell), the epithelial cell line NHBE and A549 overexpressing the angiotensin receptor ACE2 (A549\_ACE2).



**Figure 3.** Pathway enrichment of the PI, IN and DE gene sets. For each biological system (in vitro and in vivo), significantly enriched Wikipathways by the three sets of genes (PI, IN, DE in the columns) are shown (rows). The number of samples that enriched specific pathways are marked with different colours (values). Furthermore, the enriched pathways have been grouped according to more generic biological processes (cell differentiation, cell metabolism, cell death, metabolism, immune system) or molecules and structures (mRNA, viral, vascular and cilia).

interferons are key antiviral mediators, and low levels have been described in COVID-19 patients [57].

Both IN and DE genes enriched pathways related to cell differentiation, such as lung fibrosis, Wnt pathway and ectoderm differentiation. As we already reported, COVID-19 disease shares many mediators of the lung fibrosis pathogenesis, such as *NFkB*, *IL-6*, *TGF* and *INF* [58]. Furthermore, the receptor *ACE2* is a known anti-fibrotic mediator, and lung fibrosis has already been reported subsequently to the outbreak of SARS-CoV [59], making it also a plausible long-term consequence of SARS-CoV-2 viral infection. IN genes specifically enriched the Wnt pathway, which has been linked to chronic lung pathologies, including idiopathic pulmonary fibrosis, pulmonary arterial hypertension, asthma and chronic obstructive pulmonary disease [60]. Altogether, this suggests that fibrogenic alterations in the lung can be a possible long-term effect of the COVID-19 pathogenesis, as we have already recently suggested [58].

Finally, the IN gene set enriched specific biological functions represented in neither PI nor DE. Signalling related pathways, with the exception of nuclear receptors, are only present in the IN group. This indicates the central role of the IN genes in propagating the signal from the PI initial interactors to the late effector pathways. Consistently, mRNA processing pathways are only enriched in the IN group.

Therefore, the pathway enrichment of the newly identified IN set of genes reveals specific categories that represent signalling and metabolic pathways. These intermediate pathways are filling the gap between the first interactors and the late effector pathways, as well as cell differentiation pathways, suggesting possible long-term lung tissue remodelling.

### The intermediate genes are also linked to endothelial cells dysfunction and vascular remodelling

Interestingly, the IN gene set also enriched vascular related pathways. Among them, we found VEGF signaling pathways, angiogenesis, EPO signaling and extracellular matrix related pathways. Ackermann *et al.* recently showed that lung tissue of SARS-CoV-2 infected patients presented endothelial damage and significant new vessel growth [61]. The overall modulation of vascular related pathways highlighted in the IN genes, as well as the previously described cell differentiation pathways, may be an indication of endothelial remodelling and dysfunction. Endothelial dysfunction refers to a systemic condition in which the endothelium loses its physiological properties, including the tendency to promote vasodilation, fibrinolysis and platelets aggregation [62]. Different studies already proposed the endothelium as one of the main targets of SARS-CoV-2 [63–65], furthermore increasing evidence of coagulation alterations and fibrotic lesions are currently emerging in the scientific literature [63, 66]. Therefore, the new set of IN genes further strengthens the notion that the endothelial cells play a pivotal role in the COVID-19 disease and can help in predicting long-term effects in the lung in terms of vascular remodeling and dysfunction.

We further compiled five ranked lists of intermediate genes (for each *in vitro* and *in vivo* system represented in the DE space), according to the frequency in which they occurred in the list of shortest paths identified in each biological system. To obtain a final consensus rank, we merged the lists by using the Borda method (Supplementary File S6 available online at <https://academic.oup.com/bib>).

Leucine-rich repeat kinase 2 (*LRRK2*) is the most frequently visited gene in the shortest paths identified in the gene-gene network retrieved from the UKS. This gene, which has been

extensively studied for its role in Parkinson disease [67], is known to upregulate the transcriptional activity of *NFkB* by increasing phosphorylation levels of *NFkB* inhibitor alpha (*IkBα*). Hongge *et al.* proposed that *LRRK2* has the potential to be an important target for the treatment of endothelial dysfunction [68]. Furthermore, Marker *et al.* [69] demonstrated that in HIV infection, *LRRK2* decreases the levels of the angiogenesis inhibitor *BAI1* and increases the production of pro-inflammatory cytokines and phagocytosis. Given the pivotal role of *NFkB* in the COVID-19 disease, *LRRK2* is potentially important in both acute and long-term responses.

*Cullin 3* (*CUL3*), the third gene in the rank, has a role in endothelial remodelling and angiogenesis, both in physiological and pathological conditions [70].

The Exportin 1 (*XPO1*) gene is known to modulate the activity of mothers against decapentaplegic homolog 3 (*SMAD3*), a well-established initiator of epithelial mesenchymal transition (EMT) [71]. *SMAD3* is an important downstream transcription factor of TGF- $\beta$ , which regulates the transcription of extracellular matrix components involved in cellular infection [72]. Interestingly, *XPO1*, together with *SMAD3* and TGF- $\beta$ , are strongly linked to lung fibrosis [73, 74]. Similarly, heat shock protein family A (*Hsp70*) member 4 (*HSPA4*), a chaperone protein, modulates the expression of transcription factor *TWIST1*, a master regulator of morphogenesis and epithelial mesenchymal transition [75].

The histone variant *H2AX*, a sensitive marker of DNA repair machinery, is also present among the top genes of the Borda ranking. There is evidence that it plays an important role in endothelial cell proliferation under hypoxia and, more generally, in hypoxia-induced angiogenesis.

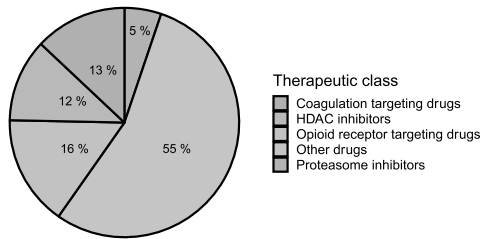
The heterogeneous nuclear ribonucleoprotein A1 (*hnRNP A1*) gene has the capability of controlling migration, proliferation and gene expression levels of vascular smooth muscle cells. A recent functional study showed that not only *hnRNP A1* is an important regulator in vascular smooth muscle cells function and lesion-induced vessel remodeling but may also represent a potential therapeutic target [76].

Finally, during lung epithelium infection, an important role in activating both the innate and adaptive immune system and the tissue repair mechanisms is also played by the estrogen receptors [77]. Furthermore, anti-inflammatory effects of estrogens have already been reported [78] and are also supported by our results since both estrogen receptors *ESR1* and *ESR2* are contained in the top ranked genes (Supplementary File S6 available online at <https://academic.oup.com/bib>). Based on our results, our novel UKS is able to highlight key genes involved in possible long-term effects of SARS-CoV-2, which are associated with vascular remodelling and endothelial dysfunction, and in some cases have already been pointed out as interesting therapeutic targets.

### Drugs targeting genes in all gene sets suggest repositioning of drugs with anti-angiogenic and immuno-modulatory properties

Given the functional importance of the IN gene set, we further investigated whether these genes could also be molecular targets of known drugs. We retrieved information about drugs targeting the PI, IN and DE gene sets from the UKS.

Highlighting drugs, which can simultaneously target multiple components of the host response (PI, IN and DE gene sets) allows to uncover possible therapeutic strategies, which can more effectively reduce the clinical consequences of the viral infection [79, 80]. We hence identified 77 drugs targeting genes



**Figure 4.** Overview of the 77 drugs targeting genes in all gene sets (PI, IN, DE). The list of 77 drugs sharing at least one target in each set of genes (PI, IN, DE) belong to four main therapeutic classes showing both immunomodulatory and anti-angiogenic properties: HDAC inhibitors (12%), proteasome inhibitors (5%), drugs targeting the opioids receptors (16%) and the coagulation cascade (13%).

in all the three gene sets of interest (Figure 4 and Supplementary File S7 available online at <https://academic.oup.com/bib>).

Among the 77 drugs, four different therapeutic classes were strongly represented: HDAC inhibitors, proteasome inhibitors, drugs targeting the coagulation cascade and drugs targeting the opioids receptors (Figure 4).

Moreover, we found cough suppressants, such as dextromethorphan, hydrocodone and pentoxifyverine, as well as expectorants and bronchodilators, such as theophylline, aminophylline and oxtriphylline [81]. These drugs are all centrally acting agents, thus exerting their effect on the lungs by inhibiting the cough centre in the brain.

Other well-represented drug categories were analgesics, antipsychotics and opioid antagonists. Haloperidol, amitriptyline, pentazocine and naltrexone, among others, belong to such categories. These drugs, together with the previously described dextromethorphan, hydrocodone and pentoxifyverine, share the same molecular targets both in the PI and IN gene sets: the sigma non-opioid intracellular receptor 1 (SIGMAR1) and the  $\mu$  opioid receptor (OPRM1), respectively (Supplementary File S7 available online at <https://academic.oup.com/bib>). Opioid drugs have a well-recognized effect on immune cells both modulating the immune system and exerting anti-inflammatory properties [82]. Besides, existing literature suggests that opioids might be able to interact with viral receptors, viral proteins, viral promoters and even modulate epigenetic mechanisms, such as the expression of anti-viral miRNAs [83]. In fact, dextromethorphan was already reported by Gordon *et al.*, because of its antiviral properties. On the other hand, dextromethorphan also shows immunomodulatory effects by decreasing *NF- $\kappa$ B* and the MAPK cascade genes activation in LPS-treated dendritic cells, and interfering with primary T-cell responses [84]. On the contrary, naltrexone, an antagonist of the  $\mu$  receptor, has been shown to revert the immunomodulatory action of opioids in several experimental models [85]. Since the sigma receptors have negligible affinity for naltrexone, it might be speculated that a significant part of the effect is exerted via direct binding to the opioid receptors. Taken together, these data suggest that compounds acting on the sigma opioid receptors might be involved in the innate and adaptive immunity in response to a SARS-CoV-2 infection and that they can have an effect in modulating the cytokine storm observed in the most severe and life-threatening stages of the disease.

Fostamatinib, a tyrosine kinase inhibitor, is also present in the list of identified drugs and importantly it targets LRRK2, the most commonly crossed IN genes in the shortest paths

derived from the UKS. Fostamatinib is currently used to treat autoimmune diseases and thrombocytopenia, but it has recently been proposed for COVID-19 disease treatment by Saha *et al.* [86]. Similar to fostamatinib, we retrieved several drugs targeting the coagulation cascade, such as kappadione, a vitamin K analogue, and menadione, used in hypoprothrombinemia treatment. It has already been shown that COVID-19 patients commonly show thrombocytopenia and are at risk of developing disseminated intravascular coagulation, even though the molecular mechanisms have been poorly described [87, 88]. Thrombocytopenia is usually associated with an excessive activation of platelets and of the coagulation cascade, which can be triggered upon viral infection. Indeed, viruses have the ability of altering the balance between procoagulant and anticoagulant homeostatic mechanisms, as well as to induce pathogenic processes such as endothelial dysfunction, Toll-like receptor activation and tissue factor pathway inhibitor activation [87, 89].

Noteworthy, the drugs listed in Supplementary File S7 available online at <https://academic.oup.com/bib> highlighted possible repositioning of HDAC inhibitors. HDAC inhibitors are a class of compounds that act on epigenetic regulation of gene expression by increasing the lysine acetylation of histones [90]. They have antiviral properties by controlling the virus replication cycle and exerting cytotoxic activity, but they also have immunomodulatory properties by regulating the production of cytokines as well as the activity of macrophages and dendritic cells [91, 92]. Gordon *et al.* [8] showed that the SARS-CoV-2 non-structural protein 5 (Nsp5) interacts with the histone deacetylases and proposed valproic acid as a therapeutic agent in COVID-19. Our UKS system was able to detect several HDAC inhibitors, which target genes in all the PI, IN and DE sets: romidepsin, belinostat, entinostat, tacedinaline, fimepinostat, panobinostat, Cudc-101 and the valproic acid itself. Specifically, the eight HDAC inhibitors targeted the HDAC2 gene present in the PI set, and the HDAC5, HDAC7 and HDAC11 present in the IN gene list, and HDAC9, HDAC1, HDAC10, HDAC3, HDAC6 and HDAC8 in the DE set. HDAC2 is a class I inhibitor located in the nucleus of the cell, where it can modulate inflammation in macrophages and monocytes by inhibiting the *NF $\kappa$ B* complex [93]. On the contrary, HDAC5 is a class II inhibitor, which can migrate into the nucleus upon phosphorylation and mediate important anti-inflammatory functions [94]. Thalidomide and its derivatives, pomalidomide and lenalidomide, also share HDAC2 as a molecular target. Thalidomide is an immunomodulatory agent and works by a number of mechanisms including the stimulation of T cells as well as decreasing TNF production. Importantly, these compounds also share anti-angiogenic properties and inhibit the proliferation of endothelial vascular cells [95].

Moreover, we identified proteasome inhibitors, sharing both antiviral and anti-angiogenic activity [96]. The ubiquitin-proteasome system plays an important role in virus replication and cell cycle, thus inhibiting virus entry, genome replication and viral protein synthesis. Proteasome inhibitors have already been pointed out as therapeutic strategies against other coronaviruses, since they can also limit the cytokine storm associated with the abnormal immunological response induced by the virus [97]. Most proteasome inhibitors can inhibit the *NF $\kappa$ B*-mediated production of IL-6, and, by inhibiting the *NF $\kappa$ B* transcription factor, they also exert an important anti-angiogenic effect [98]. Remarkably, HDAC inhibitors, proteasome inhibitors and thalidomide derivatives, are all currently used as a therapeutic regimen against multiple myeloma, an oncological condition in which myeloma cells produce a microenvironment

enriched with pro-angiogenic factors, such as VEGF and IL-6 [95]. In conclusion, the four classes of drugs identified by the UKS share both immuno-modulatory and anti-angiogenic properties and are therefore good candidates in counteracting both the acute cytokine storm as well as endothelial and vascular complications.

## Conclusions

Characterizing the cascade of events taking place at multiple levels in response to SARS-CoV-2 infection is urgently needed as the COVID-19 pandemic keeps rampaging worldwide. Here, we interrogated a unified network of public biomedical data, the Unified Knowledge Space (UKS), in order to elucidate the molecular alterations characterizing the SARS-CoV-2 infection.

By assuming that early viral responses are mediated by virus-interacting genes, while the downstream effects of infection are mediated by genes whose expression is altered, we interrogated the UKS in search of a novel set of intermediate genes that would help to further characterize the COVID-19 pathogenesis. Our analysis highlighted genes representing functions related to fibrosis and vascular remodelling, implying further long-term consequences of SARS-CoV-2 infection. Furthermore, we identified a set of drugs with at least one target present in each of the identified gene sets: proteins known to interact with SARS-CoV-2 (PI, as defined by Gordon et al. [8]), differentially expressed (DE) genes in multiple biological systems infected by SARS-CoV-2 (Blanco-Melo et al. [9]) and intermediate genes (IN, newly discovered here). Our results point to therapeutic classes with immunomodulatory and anti-angiogenic roles.

In conclusion, the robust network-based approach applied here helps to shed light on the details of the SARS-CoV-2-host interaction, suggesting possible long-term effects of the viral infections, and highlights important therapeutic targets, paving the way to new drug repositioning studies. Furthermore, due to the high flexibility of the UKS, our strategy can be applied to study the molecular alterations induced by other diseases or by the exposure to drugs or chemicals.

### Key Points

- Integrated molecular network analysis can help to clarify the pathogenesis of complex diseases and suggest novel drug targets.
- By mapping SARS-CoV-2 first physical interactors and COVID-19 downstream differentially expressed genes on the integrated human molecular network, we identified a new set of intermediate genes.
- The newly discovered set of intermediate genes underlies important aspects of COVID-19 pathogenesis and long-term consequences, pointing to lung tissue remodelling and fibrosis.
- We highlighted immuno-modulatory and anti-angiogenic drugs targeting multiple genes in each and every relevant set: physical interactors, intermediate and downstream effectors.

## Supplementary data

Supplementary data are available online at *Briefings in Bioinformatics*.

## Funding

Academy of Finland (grant number 322761).

## Conflict of interest

The authors declare no conflict of interest.

## Author contributions

A.P. collected, integrated and analysed the raw data, developed the computational framework, drafted the manuscript. G.d.G. analysed and interpreted the results, drafted the manuscript. A.F. analysed the transcriptomics data and revised the manuscript. A.D.L. and P.A.S.K. interpreted the results and drafted the manuscript. A.S. supervised the development of the computational framework and revised the manuscript. D.G. conceived and supervised the project, edited the manuscript.

## References

1. Girija ASS, Shankar EM, Larsson M. Could SARS-CoV-2-induced hyperinflammation magnify the severity of coronavirus disease (COVID-19) leading to acute respiratory distress syndrome? *Front Immunol* 2020;11:1206.
2. Yang X, Yu Y, Xu J, et al. Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. *Lancet Respir Med* 2020;8:475–81.
3. O'Neill LAJ, Netea MG. BCG-induced trained immunity: can it offer protection against COVID-19? *Nat Rev Immunol* 2020;20:335–7.
4. Zhu S, Guo X, Geary K, et al. Emerging therapeutic strategies for COVID-19 patients. *Discoveries (Craiova)* 2020;8:e105.
5. Jeong GU, Song H, Yoon GY, et al. Therapeutic strategies against COVID-19 and structural characterization of SARS-CoV-2: a review. *Front Microbiol* 2020;11:1723.
6. Matricardi PM, Dal Negro RW, Nisini R. The first, holistic immunological model of COVID-19: implications for prevention, diagnosis, and public health measures. *Pediatr Allergy Immunol* 2020;31(5):454–70.
7. Varga Z, Flammer AJ, Steiger P, et al. Endothelial cell infection and endothelitis in COVID-19. *Lancet* 2020;395:1417–8.
8. Gordon DE, Jang GM, Bouhaddou M, et al. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* 2020;583:459–68.
9. Blanco-Melo D, Nilsson-Payant BE, Liu W-C, et al. Imbalanced host response to SARS-CoV-2 drives development of COVID-19. *Cell* 2020;181:1036.e9–45.
10. Wang X, Williams C, Liu ZH, et al. Big data management challenges in health research—a literature review. *Brief Bioinform* 2019;20:156–67.
11. Manzoni C, Kia DA, Vandrovicova J, et al. Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. *Brief Bioinform* 2018;19:286–302.
12. Tadist K, Najah S, Nikolov NS, et al. Feature selection methods and genomic big data: a systematic review. *J Big Data* 2019;6:79.

13. Liang X, Li D, Song M, et al. Predicting biomedical relationships using the knowledge and graph embedding cascade model. *PLoS One* 2019;**14**:e0218264.
14. Nicholson DN, Greene CS. Constructing knowledge graphs and their biomedical applications. *Comput Struct Biotechnol J* 2020;**18**:1414–28.
15. Nickel M, Murphy K, Tresp V, et al. A review of relational machine learning for knowledge graphs. *Proc IEEE* 2016;**104**:11–33.
16. Ehrlinger L, Wöfl W. Towards a definition of knowledge graphs. *SEMANTiCS 2016: 12th International Conference on Semantic Systems*, Leipzig, Germany, New York (NY), United States: Association for Computing Machinery, 2016. p. 1695
17. Cheng F, Liu C, Jiang J, et al. Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS Comput Biol* 2012;**8**:e1002503.
18. Sosa DN, Derry A, Guo M, et al. A literature-based knowledge graph embedding method for identifying drug repurposing opportunities in rare diseases. *Pac Symp Biocomput* 2020;**25**:463–74.
19. Shen Z, Zhang Y-H, Han K, et al. miRNA-disease association prediction with collaborative matrix factorization. *Complexity* 2017;**2017**:1–9.
20. Goh K-I, Cusick ME, Valle D, et al. The human disease network. *Proc Natl Acad Sci U S A* 2007;**104**:8685–90.
21. Gligorijević V, Pržulj N. Methods for biological data integration: perspectives and challenges. *J R Soc Interface* 2015; **12**.
22. Zhang X, Zhang R, Jiang Y, et al. The expanded human disease network combining protein-protein interaction information. *Eur J Hum Genet* 2011;**19**:783–8.
23. Davis DA, Chawla NV. Exploring and exploiting disease interactions from multi-relational gene and phenotype networks. *PLoS One* 2011;**6**:e22670.
24. Kivela M, Arenas A, Barthelemy M, et al. Multilayer networks. *J Complex Netw* 2014;**2**:203–71.
25. Boccaletti S, Bianconi G, Criado R, et al. The structure and dynamics of multilayer networks. *Phys Rep* 2014;**544**: 1–122.
26. Hunt SE, McLaren W, Gil L, et al. Ensembl variation resources. *Database (Oxford)* 2018;**2018**.
27. Tarazona S, Garcia-Alcalde F, Dopazo J, et al. Differential expression in RNA-seq: a matter of depth. *Genome Res* 2011;**21**:2213–23.
28. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**:550.
29. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B Methodol* 1995;**57**:289–300.
30. 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**: D408–14.
31. López Y, Nakai K, Patil A. HitPredict version 4: comprehensive reliability scoring of physical protein-protein interactions from more than 100 species. *Database (Oxford)* 2015;**2015**.
32. Patil A, Nakai K, Nakamura H. HitPredict: a database of quality assessed protein-protein interactions in nine species. *Nucleic Acids Res* 2011;**39**:D744–9.
33. Kanehisa M, Furumichi M, Tanabe M, et al. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* 2017;**45**:D353–61.
34. Szklarczyk D, Gable AL, Lyon D, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* 2019;**47**: D607–13.
35. Chiang T, Scholtens D, Sarkar D, et al. Coverage and error models of protein-protein interaction data by directed graph analysis. *Genome Biol* 2007;**8**:R186.
36. Huang H, Bader JS. Precision and recall estimates for two-hybrid screens. *Bioinformatics* 2009;**25**:372–8.
37. Huang H, Jedynak BM, Bader JS. Where have all the interactions gone? Estimating the coverage of two-hybrid protein interaction maps. *PLoS Comput Biol* 2007;**3**: e214.
38. Martha V-S, Liu Z, Guo L, et al. Constructing a robust protein-protein interaction network by integrating multiple public databases. *BMC Bioinformatics* 2011;**12**(Suppl 10): S7.
39. Wishart DS, Feunang YD, Guo AC, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* 2018;**46**:D1074–82.
40. Koscielny G, An P, Carvalho-Silva D, et al. Open targets: a platform for therapeutic target identification and validation. *Nucleic Acids Res* 2017;**45**:D985–94.
41. Kim S, Chen J, Cheng T, et al. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res* 2019;**47**:D1102–9.
42. Xin J, Mark A, Afrasiabi C, et al. High-performance web services for querying gene and variant annotation. *Genome Biol* 2016;**17**:91.
43. Wu C, Macleod I, Su AI. BioGPS and MyGene.info: organizing online, gene-centric information. *Nucleic Acids Res* 2013;**41**:D561–5.
44. Hagberg AA, Schult DA, Swart PJ. Exploring network structure, dynamics, and function using networkX. *Proceedings of the 7th Python in Science Conference (SciPy2008)*, Pasadena, CA, USA, 2008.
45. Dijkstra EW. A note on two problems in connection with graphs. *Numer Math* 1959;**1**:269–71.
46. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* 2020;**17**:261–72.
47. Seabold S, Perktold J. Statsmodels: econometric and statistical modeling with Python. *Proceedings of the 9th Python in Science Conference*, Austin, Texas, 2010. pp. 92–6
48. Chen EY, Tan CM, Kou Y, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* 2013;**14**:128.
49. Kuleshov MV, Jones MR, Rouillard AD, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* 2016;**44**:W90–7.
50. Scala G, Serra A, Marwah VS, et al. FunMappOne: a tool to hierarchically organize and visually navigate functional gene annotations in multiple experiments. *BMC Bioinformatics* 2019;**20**:79.
51. Schimek MG, Budinská E, Kugler KG, et al. TopKLists: a comprehensive R package for statistical inference, stochastic aggregation, and visualization of multiple omics ranked lists. *Stat Appl Genet Mol Biol* 2015;**14**:311–6.
52. Du Z-P, Wu B-L, Wang S-H, et al. Shortest path analyses in the protein-protein interaction network of NGAL (neutrophil gelatinase-associated lipocalin) overexpression in esophageal squamous cell carcinoma. *Asian Pac J Cancer Prev* 2014;**15**:6899–904.

53. Simões SN, Martins-Jr DC, Brentani H, et al. Shortest paths ranking methodology to identify alterations in PPI networks of complex diseases. *BCB'16: ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, New York (NY), United States: Association for Computing Machinery, 2012. pp. 561–3.
54. Ren Y, Ay A, Kahveci T. Shortest path counting in probabilistic biological networks. *BMC Bioinformatics* 2018;**19**:465.
55. M A, Vazquez-Calvo A, Caridi F, et al. Lipid involvement in viral infections: present and future perspectives for the design of antiviral strategies. *Lipid Metabolism* 2013;**291**–322. doi: 10.5772/51068.
56. Vabret N, Britton GJ, Gruber C, et al. Immunology of COVID-19: current state of the science. *Immunity* 2020;**52**: 910–41.
57. Acharya D, Liu G, Gack MU. Dysregulation of type I interferon responses in COVID-19. *Nat Rev Immunol* 2020;**20**:397–8.
58. Kinaret PAS, Del Giudice G, Greco D. Covid-19 acute responses and possible long term consequences: what nanotoxicology can teach us. *Nano Today* 2020;**35**:100945.
59. Zuo W, Zhao X, Chen Y-G. SARS coronavirus and lung fibrosis. *Molecular biology of the SARS-Coronavirus*. 2010. pp. 247–58.
60. Baarsma HA, Königshoff M. “WNT-er is coming”: WNT signalling in chronic lung diseases. *Thorax* 2017;**72**:746–59.
61. Ackermann M, Verleden SE, Kuehnel M, et al. Pulmonary vascular endothelialitis, thrombosis, and angiogenesis in Covid-19. *N Engl J Med* 2020;**383**:120–8.
62. Hadi HAR, Carr CS, Al Suwaidi J. Endothelial dysfunction: cardiovascular risk factors, therapy, and outcome. *Vasc Health Risk Manag* 2005;**1**:183–98.
63. Lemke G, Silverman GJ. Blood clots and TAM receptor signalling in COVID-19 pathogenesis. *Nat Rev Immunol* 2020;**20**(7):395–6.
64. Sardu C, Gambardella J, Morelli MB, et al. Hypertension, Thrombosis, Kidney failure, and diabetes: is COVID-19 an Endothelial disease? A comprehensive evaluation of clinical and basic evidence, St. Alban-Anlage 66, 4052 Basel, Switzerland. 2020. doi: 10.3390/jcm9051417.
65. Huertas A, Montani D, Savale L, et al. Endothelial cell dysfunction: a major player in SARS-CoV-2 infection (COVID-19)? *Eur Respir J* 2020;**56**(1):2001634.
66. George PM, Wells AU, Jenkins RG. Pulmonary fibrosis and COVID-19: the potential role for antifibrotic therapy. *Lancet Respir Med* 2020;**8**:807–15.
67. Li J-Q, Tan L, Yu J-T. The role of the LRRK2 gene in parkinsonism. *Mol Neurodegener* 2014;**9**:47.
68. Hongge L, Xexin G, Xiaojie M, et al. The role of LRRK2 in the regulation of monocyte adhesion to endothelial cells. *J Mol Neurosci* 2015;**55**:233–9.
69. Marker DF, Puccini JM, Mockus TE, et al. LRRK2 kinase inhibition prevents pathological microglial phagocytosis in response to HIV-1 Tat protein. *J Neuroinflammation* 2012;**9**:261.
70. Sakaue T, Maekawa M, Nakayama H, et al. Prospect of divergent roles for the CUL3 system in vascular endothelial cell function and angiogenesis. *J Biochem* 2017;**162**:237–45.
71. Pan X, Wang B, Yuan T, et al. Analysis of combined transcriptomes identifies gene modules that differentially respond to pathogenic stimulation of vascular smooth muscle and endothelial cells. *Sci Rep* 2018;**8**:395.
72. Gough NR. Enhancing and inhibiting TGF-signaling in infection. *Sci Signal* 2015;**8**:ec9–9.
73. Walton KL, Johnson KE, Harrison CA. Targeting TGF- $\beta$  mediated SMAD signaling for the prevention of fibrosis. *Front Pharmacol* 2017;**8**:461.
74. Bruccoleri A, Rondinone O, Andriani F, et al. Abstract 1912: inhibition of Exportin-1 function reverses the protumorigenic potential of lung fibrotic microenvironments. *Tumor Biol* 2019;**79**(13):1912–2.
75. Park JM, Kim JW, Hahm KB. HSPA4, the “evil chaperone” of the HSP family, delays gastric ulcer healing. *Dig Dis Sci* 2015;**60**:824–6.
76. Zhang L, Chen Q, An W, et al. Novel pathological role of hnRNPA1 (heterogeneous nuclear ribonucleoprotein A1) in vascular smooth muscle cell function and Neointima hyperplasia. *Arterioscler Thromb Vasc Biol* 2017;**37**:2182–94.
77. Suba Z. Prevention and therapy of COVID-19 via exogenous estrogen treatment for both male and female patients. *J Pharm Pharm Sci* 2020;**23**:75–85.
78. Kim KH, Young BD, Bender JR. Endothelial estrogen receptor isoforms and cardiovascular disease. *Mol Cell Endocrinol* 2014;**389**:65–70.
79. Altay O, Mohammadi E, Lam S, et al. Current status of COVID-19 therapies and drug repositioning applications. *iScience* 2020;**23**:101303.
80. Li L, Li R, Wu Z, et al. Therapeutic strategies for critically ill patients with COVID-19. *Ann Intensive Care* 2020;**10**:45.
81. Zhou C, Gao C, Xie Y, et al. COVID-19 with spontaneous pneumomediastinum. *Lancet Infect Dis* 2020;**20**:510.
82. Franchi S, Moschetti G, Amodeo G, et al. Do all opioid drugs share the same immunomodulatory properties? A review from animal and human studies. *Front Immunol* 2019;**10**: 2914.
83. Tahamtan A, Tavakoli-Yaraki M, Mokhtari-Azad T, et al. Opioids and viral infections: a double-edged sword. *Front Microbiol* 2016;**7**:970.
84. Chen D-Y, Song P-S, Hong J-S, et al. Dextromethorphan inhibits activations and functions in dendritic cells. *Clin Dev Immunol* 2013;**2013**:125643.
85. Rousseaux CG, Greene SF. Sigma receptors [ $\sigma$ Rs]: biology in normal and diseased states. *J Recept Signal Transduct Res* 2015;**36**(4):1–62.
86. Saha S, Halder AK, Bandyopadhyay SS, et al. Is Fostamatinib a possible drug for COVID-19? – A computational study. 2020. doi: 10.31219/osf.io/7hgpj.
87. Giannis D, Ziogas IA, Gianni P. Coagulation disorders in coronavirus infected patients: COVID-19, SARS-CoV-1, MERS-CoV and lessons from the past. *J Clin Virol* 2020;**127**: 104362.
88. Boccia M, Aronne L, Celia B, et al. COVID-19 and coagulative axis: review of emerging aspects in a novel disease. *Monaldi Arch Chest Dis* 2020;**90**.
89. Subramaniam S, Scharrer I. Procoagulant activity during viral infections. *Front Biosci (Landmark Ed)* 2018;**23**: 1060–81.
90. Eckschlager T, Plch J, Stiborova M, et al. Histone deacetylase inhibitors as anticancer drugs. *Int J Mol Sci* 2017;**18**: 1414.
91. Herbein G, Wendling D. Histone deacetylases in viral infections. *Clin Epigenetics* 2010;**1**:13–24.
92. Schotterl S, Brennenstuhl H, Naumann U. Modulation of immune responses by histone deacetylase inhibitors. *Crit Rev Oncog* 2015;**20**:139–54.
93. Ito K, Hanazawa T, Tomita K, et al. Oxidative stress reduces histone deacetylase 2 activity and enhances IL-8 gene

- expression: role of tyrosine nitration. *Biochem Biophys Res Commun* 2004;**315**:240–5.
94. Fuchikami M, Yamamoto S, Morinobu S, et al. The potential use of histone deacetylase inhibitors in the treatment of depression. *Prog Neuro-Psychopharmacol Biol Psychiatry* 2016;**64**:320–4.
  95. Giuliani N, Storti P, Bolzoni M, et al. Angiogenesis and multiple myeloma. *Cancer Microenviron* 2011;**4**:325–37.
  96. Schneider SM, Pritchard SM, Wudiri GA, et al. Early steps in herpes simplex virus infection blocked by a proteasome inhibitor. *MBio* 2019;**10**(3):e00732-19.
  97. Longhitano L, Tibullo D, Giallongo C, et al. Proteasome inhibitors as a possible therapy for SARS-CoV-2. *Int J Mol Sci* 2020;**21**:3622.
  98. Almond JB, Cohen GM. The proteasome: a novel target for cancer chemotherapy. *Leukemia* 2002;**16**:433–43.



# PUBLICATION II

## **VOLTA: adVanced mOLecular neTwork Analysis**

Alisa Pavel, Antonio Federico, Giusy del Giudice, Angela Serra, Dario Greco

Bioinformatics, Volume 37, Issue 23, December 2021, Pages 4587–4588  
<https://doi.org/10.1093/bioinformatics/btab642>

**Publication is licensed under a Creative Commons Attribution 4.0  
International License CC-BY-NC-ND**



Systems biology

# VOLTA: adVanced mOLecular neTwork Analysis

Alisa Pavel<sup>1,2,3</sup>, Antonio Federico<sup>1,2,3</sup>, Giusy del Giudice<sup>1,2,3</sup>, Angela Serra<sup>1,2,3</sup> and Dario Greco <sup>1,2,3,4,\*</sup>

<sup>1</sup>Faculty of Medicine and Health Technology, Tampere University, Tampere 33520, Finland, <sup>2</sup>BioMediTech Institute, Tampere University, Tampere 33520, Finland, <sup>3</sup>Finnish Hub for Development and Validation of Integrated Approaches (FHAIVE), Faculty of Medicine and Health Technology, Tampere University, Tampere 33520, Finland and <sup>4</sup>Institute of Biotechnology, University of Helsinki, Helsinki 00790, Finland

\*To whom correspondence should be addressed.

Associate Editor: Anthony Mathelier

Received on May 3, 2021; revised on August 11, 2021; editorial decision on September 2, 2021; accepted on September 5, 2021

## Abstract

**Motivation:** Network analysis is a powerful approach to investigate biological systems. It is often applied to study gene co-expression patterns derived from transcriptomics experiments. Even though co-expression analysis is widely used, there is still a lack of tools that are open and customizable on the basis of different network types and analysis scenarios (e.g. through function accessibility), but are also suitable for novice users by providing complete analysis pipelines.

**Results:** We developed VOLTA, a Python package suited for complex co-expression network analysis. VOLTA is designed to allow users direct access to the individual functions, while they are also provided with complete analysis pipelines. Moreover, VOLTA offers when possible multiple algorithms applicable to each analytical step (e.g. multiple community detection or clustering algorithms are provided), hence providing the user with the possibility to perform analysis tailored to their needs. This makes VOLTA highly suitable for experienced users who wish to build their own analysis pipelines for a wide range of networks as well as for novice users for which a ‘plug and play’ system is provided.

**Availability and implementation:** The package and used data are available at GitHub: <https://github.com/fhaive/VOLTA> and [10.5281/zenodo.5171719](https://zenodo.org/record/5171719).

**Contact:** [dario.greco@tuni.fi](mailto:dario.greco@tuni.fi)

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

## 1 Introduction

Co-expression network analysis has become popular to characterize gene–gene expression patterns from omics data by providing insight into the differential gene co-expression patterns and their local and global organizations, between different biological conditions (van Dam *et al.*, 2018; Liu *et al.*, 2017). Currently three main classes of network analysis software exist to (i) infer co-expression networks from experimental data (Marwah *et al.*, 2018), (ii) investigate the properties of individual networks (Hagberg *et al.*, 2008) and (iii) compare multiple networks (Proost and Mutwil, 2018), while flexible, comprehensive tools are currently still missing. We therefore developed VOLTA, a Python package that combines traditional network metrics with functions adjusted to the comparison and evaluation of co-expression networks. In addition, VOLTA is highly versatile by nature, allowing users easy access to all functionalities and parameters. This helps the users to create analytical pipelines to answer a wide range of biological questions, which is in contrast to many other available software/tools which are restricted to

specific steps through their implementation (Proost and Mutwil, 2018; Supplementary Text). To the best of our knowledge there is currently no other package available, which combines a diverse set of network analysis methods into a single package, completely exposes its internal functionalities and therefore is highly versatile.

## 2 Implementation

VOLTA consists of seven modules (Supplementary Fig. S2), which can be used independently or in combination to create complex analytical pipelines. VOLTA is implemented in Python 3 and allows users deep access to all functionalities and parameter settings. In addition to the main function modules, VOLTA provides six predefined pipeline wrappers. Three fully functional pipelines are provided in the form of Jupyter Notebooks respectively addressing: (i) *clustering of multiple networks* employing global and local similarities; (ii) *identification of common connectivity patterns in a set of*

networks and (iii) *network–network comparison* based on their nodes, edges and communities (Supplementary Text S1).

### 3 Application

To demonstrate the functionalities and applicability of VOLTA in co-expression network analysis, we selected three possible analysis scenarios. The networks for this study were generated from the Lincs 1000 data (Supplementary Text S4.1). In the first case, in order to describe the transcriptional perturbation induced on A549 cells by treatment with dasatinib and mitoxantrone, we compared the characteristics (i.e. connectivity) of the two co-expression networks by exploiting the functionalities of the VOLTA package. Such an analysis allowed the characterization of the specific mechanism of action of the considered chemotherapeutic drugs. Evaluation of difference in gene centrality in the two networks, showcases a high difference in centrality among the networks of *OXA1L*, *YME1L1* and *DNAJC15* genes, suggesting an involvement of mitoxantrone in the impairment of mitochondrial function, as has been previously demonstrated (Rossato *et al.*, 2014). Comparison of pathway enrichment of the modules of the two networks showcases the difference in mechanisms of mitoxantrone and dasatinib. Modules detected in the mitoxantrone network enrich for DNA double strand break pathways, highlighting the genotoxic effect of mitoxantrone. On the other hand, functional characterization of the modules in the dasatinib network highlight the involvement in the intracellular signaling processes (Supplementary Text S4.2).

In the second case study, we aimed to assess the impact of the different molecular makeup of 20 cancer cell lines on the mechanism of action of dasatinib (Supplementary Text S4). Exploiting VOLTA functionalities for this aim allowed the investigation of drug sensitivity profiles of cancer cell lines to dasatinib treatment and to identify clusters of similarly responding cell lines. The three clusters that could be identified were (i) a cluster mainly made up of breast (cancer) related tissues, (ii) one containing ‘normal’ samples from different tissues and (iii) another one containing different tissue types—not fitting into the previous two clusters (Supplementary Table S8). In the third analysis, we showcased and characterized the statistical sub-graph of the breast related tissue cluster. Investigation of the cluster characterized sub-graph reveals genes that are involved in processes related to cell cycle, differentiation and metabolism as central. Pathway enrichment of the modules of the characterized sub-graph indicates a deregulation of immune-related pathways, together with cell cycle and DNA repair machinery (Supplementary Table S10).

### 4 Discussion

To date, many network analysis software solutions have been proposed, which have often either very general purpose (Hagberg *et al.*, 2008) or they are specialized packages to solve a specific problem (Rossetti *et al.*, 2019). Software solutions for co-expression network analysis, on the other hand, are commonly optimized for a single analysis pipeline or step (Marwah *et al.*, 2018; Proost and Mutwil, 2018). While these tools are easy to use, they can have the downside of being non-adaptable to other problems. This can for example result through stringent input format requirements, or commonly that

individual functionalities are implemented in such a way that they are not accessible from outside the provided software, which often means that individual functionalities (of a pipeline) cannot be re-used outside the ‘intended’ flow as well as that parameter adjustment is restricted (Supplementary Text S2). We therefore developed VOLTA, which combines a diverse set of exposed functions, applicable in many different fields of network analysis and aims, when possible, to provide different algorithms for a given task (for example a diverse set of community detection algorithms is provided). This allows users to customize their pipelines, for example based on their network structure, or allows the application of ensemble methods. In addition, pipelines (which can easily be modified by users due to being provided as Jupyter Notebook files (<https://github.com/fhaive/VOLTA/tree/master/jupyternotebooks>)) for specific analysis in the domain of co-expression networks are provided. This allows inexperienced users a plug-and-play experience, while more advanced users have the possibility to construct customized pipelines.

### 5 Conclusion

Here, we presented VOLTA, a Python package highly adapted to biological network analysis (with a focus on co-expression networks). It is the first package providing a wide range of functionalities adaptable to different studies in Python, which is both suited to naive as well as expert users. The usability and applicability of VOLTA in (co-expression) network analysis has been highlighted in the performed case studies.

### Acknowledgement

The authors thank Troy Faithfull for his comments on the manuscript.

### Funding

This study was supported by Academy of Finland [322761].

*Conflict of Interest:* none declared.

### References

- van Dam, S. *et al.* (2018) Gene co-expression analysis for functional classification and gene-disease predictions. *Brief. Bioinf.*, **19**, 575–592.
- Hagberg, A.A. *et al.* (2008) Exploring network structure, dynamics, and function using networkx. <https://www.osti.gov/biblio/960616-exploring-network-structure-dynamics-function-using-networkx>.
- Liu, W. *et al.* (2017) Weighted gene co-expression network analysis in biomedicine research. *Sheng Wu Gong Cheng Xue Bao*, **33**, 1791–1801. [].
- Marwah, V.S. *et al.* (2018) Inform: inference of network response modules. *Bioinformatics*, **34**, 2136–2138.
- Proost, S. and Mutwil, M. (2018) CoNekT: an open-source framework for comparative genomic and transcriptomic network analyses. *Nucleic Acids Res.*, **46**, W133–W140.
- Rossato, L.G. *et al.* (2014) Mitochondrial cumulative damage induced by mitoxantrone: late onset cardiac energetic impairment. *Cardiovasc. Toxicol.*, **14**, 30–40.
- Rossetti, G. *et al.* (2019) CDLIB: a python library to extract, compare and evaluate communities from complex networks. *Appl. Netw. Sci.*, **4**, 52.

# PUBLICATION III

## **KNeMAP: a network mapping approach for knowledge-driven comparison of transcriptomic profiles**

Alisa Pavel, Giusy del Giudice, Michele Fratello, Leo Ghemtio, Antonio Di Lieto, Jari Yli-Kauhaluoma, Henri Xhaard, Antonio Federico, Angela Serra, Dario Greco



Bioinformatics, Volume 39, Issue 6, June 2023  
<https://doi.org/10.1093/bioinformatics/btad341>

**Publication is licensed under a Creative Commons Attribution 4.0  
International License CC-BY-NC-ND**



## Gene expression

# KNeMAP: a network mapping approach for knowledge-driven comparison of transcriptomic profiles

Alisa Pavel <sup>1,2,3</sup>, Giusy del Giudice<sup>1,2,3</sup>, Michele Fratello<sup>1,2,3</sup>, Leo Ghemtio<sup>4</sup>, Antonio Di Lieto<sup>5</sup>, Jari Yli-Kauhaluoma<sup>4</sup>, Henri Xhaard<sup>4</sup>, Antonio Federico<sup>1,2,3,6</sup>, Angela Serra<sup>1,2,3,6</sup>, Dario Greco <sup>1,2,3,7,8,\*</sup>

<sup>1</sup>Faculty of Medicine and Health Technology, Tampere University, 33520 Tampere, Finland

<sup>2</sup>BioMediTech Institute, Tampere University, 33520 Tampere, Finland

<sup>3</sup>Finnish Hub for Development and Validation of Integrated Approaches (FHAIVE), 33520 Tampere, Finland

<sup>4</sup>Drug Research Program, Division of Pharmaceutical Biosciences, Faculty of Pharmacy, University of Helsinki, 00790 Helsinki, Finland

<sup>5</sup>Mental Health Services, Landspítali University Hospital, 101 Reykjavík, Iceland

<sup>6</sup>Tampere Institute for Advanced Study, 33520 Tampere, Finland

<sup>7</sup>Institute of Biotechnology, University of Helsinki, 00790 Helsinki, Finland

<sup>8</sup>Division of Pharmaceutical Biosciences, Faculty of Pharmacy, University of Helsinki, 00790 Helsinki, Finland

\*Corresponding author. Faculty of Medicine and Health Technology, Tampere University, Arvo Ylpön katu 34, 33520 Tampere, Finland.

E-mail: dario.greco@tuni.fi

Associate Editor: Alfonso Valencia

## Abstract

**Motivation:** Transcriptomic data can be used to describe the mechanism of action (MOA) of a chemical compound. However, omics data tend to be complex and prone to noise, making the comparison of different datasets challenging. Often, transcriptomic profiles are compared at the level of individual gene expression values, or sets of differentially expressed genes. Such approaches can suffer from underlying technical and biological variance, such as the biological system exposed on or the machine/method used to measure gene expression data, technical errors and further neglect the relationships between the genes. We propose a network mapping approach for knowledge-driven comparison of transcriptomic profiles (KNeMAP), which combines genes into similarity groups based on multiple levels of prior information, hence adding a higher-level view onto the individual gene view. When comparing KNeMAP with fold change (expression) based and deregulated gene set-based methods, KNeMAP was able to group compounds with higher accuracy with respect to prior information as well as is less prone to noise corrupted data.

**Result:** We applied KNeMAP to analyze the Connectivity Map dataset, where the gene expression changes of three cell lines were analyzed after treatment with 676 drugs as well as the Fortino *et al.* dataset where two cell lines with 31 nanomaterials were analyzed. Although the expression profiles across the biological systems are highly different, KNeMAP was able to identify sets of compounds that induce similar molecular responses when exposed on the same biological system.

**Availability and implementation:** Relevant data and the KNeMAP function is available at: <https://github.com/fhaive/KNeMAP> and 10.5281/zenodo.7334711.

## 1 Introduction

A fundamental challenge in compound safety and efficacy assessment is to understand the multi-scale mechanistic effects that compounds have on genes, cells, tissues, and organisms. Toxicogenomics approaches can be used to characterize the mechanism of action (MOA) of a compound (Gao *et al.* 2021), through the use of transcriptomics (Federico *et al.* 2020, Kinaret *et al.* 2020b, Serra *et al.* 2020). In addition, the comparison of molecular alteration profiles allows to identify similarities between phenotypic entities and to make conclusions about possible phenotypic changes of an exposure (Kinaret *et al.* 2020b). Transcriptomics data are complex and prone to technical and biological variability and noise (Raser and O'Shea 2005, Freytag *et al.* 2015, Federico *et al.* 2020, Fratello *et al.* 2022). Therefore many variables need to be

considered when comparing expression profiles, especially coming from different datasets or (biological) systems.

Methods to compare gene expression or gene expression alteration profiles aim to analyze lists of genes ordered by their expression levels as measured by DNA microarrays or RNA sequencing (Federico *et al.* 2020, Kinaret *et al.* 2020b). A common metric used for this is the correlation (Freytag *et al.* 2015, Serra *et al.* 2018, Serra *et al.* 2020). Differential analysis or the comparison of deregulated genes is another method, where the affected genes are compared with respect to a control, instead of using the expression values directly (Marwah *et al.* 2019, Federico *et al.* 2020). In this case, the lists of deregulated genes are directly compared to highlight differences and commonalities. Alternatively their functional profiles are compared through pathway enrichment (Federico *et al.* 2020, Serra *et al.* 2022b).

Received: November 23, 2022. Revised: April 14, 2023. Editorial Decision: May 21, 2023. Accepted: May 23, 2023

© The Author(s) 2023. Published by Oxford University Press.

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

The approach suggested in this study, a network mapping approach for knowledge-driven comparison of transcriptomic profiles (KNeMAP), builds on the assumption that genes can be grouped together based on higher level classifications, such as functions, processes or evolutionary origin. Therefore the individual gene view is replaced by a “similar gene” view, where instead of considering genes individually, a set of genes are grouped together based on multi-level prior knowledge. This gene grouping is used to create a feature vector for each experimental instance, which can be used in downstream analysis, such as clustering or machine learning (ML) applications, where often a numeric feature vector is needed as input (Serra *et al.* 2020, Fratello *et al.* 2022). This is in contrast to many functional enrichment applications, where individual pathway names are returned, that cannot be directly provided as input to such downstream ML applications.

In addition since KNeMAP is prior knowledge dependent, new feature vectors can be computed for new data, without the need to re-process existing data, since the feature vectors as long as computed from the same prior knowledge are comparable between each other. For the same reason it is also possible to compare exposure fingerprints via KNeMAP across datasets. Another difference to traditional functional enrichment is that we define gene similarity as multi-view, across multiple different data layers, capturing functional, interactional, and associational gene (product) similarities.

Here, we showcase the effectiveness of the KNeMAP method by applying it on the CMap (Lamb *et al.* 2006) dataset to compare the transcriptomic profiles of drugs across three different cell lines (biological systems), as well as the Fortino *et al.* (2022) dataset to compare the transcriptomic profiles of engineered nanomaterials (ENMs) across two different human cell lines. In addition, we compare the CMap (Lamb *et al.* 2006) and Fortino *et al.* data with each other to identify for each ENM, the drug that shows the most similar transcriptomic alterations across all biological systems. We also compare our method with three other approaches based on correlation of the gene expression fold changes (in comparison to the control gene expression), gene deregulation analysis as well as a Gene Set Enrichment Analysis (GSEA)-based methodology (Subramanian *et al.* 2005, Iorio *et al.* 2010).

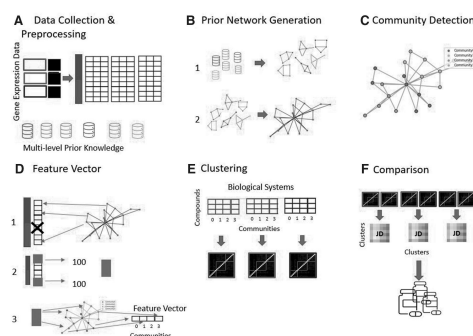
## 2 Materials and methods

### 2.1 Data collection and prior network

In order to investigate the difference between the transcriptomic alterations induced by small molecules on different biological systems, we downloaded microarray data including a set of compounds, tested on different systems (Fig. 1A), as described in Lamb *et al.* (2006) (CMap) and as described in Fortino *et al.* (2022). The processing of the data is described in the Supplementary Materials (Methods—Collection of Expression Data and Pre-Processing).

#### 2.1.1 Prior network creation and community detection

In order to build a robust gene network, we collected multiple data layers and datasets, covering different aspects of a gene’s function, relationships, and structure (Fig. 1A). By combining these data, we created a weighted network that captures multiple views of “gene similarity.” For example two genes can be considered as similar, based on their structural or ancestral similarities, on their functional similarities (e.g. takes part in



**Figure 1.** Description of the proposed methodology. (A) The collection and pre-processing of the gene expression data. The values are sorted by their  $\pm \log_{FC} * -\log_{10}(Pval)$  (FCP) values. In addition, different layers of gene (product) information are collected, such as protein family, homolog, protein-protein interaction (PPI) information as well as associations to phenotypes, compounds, and gene ontology (GO) (The Gene Ontology Consortium 2021) terms. (B) The individual gene (product) information data types are converted into gene-gene similarity networks (1). The individual networks are merged into a single weighted gene-gene similarity network, the prior network (2). (C) The prior network is partitioned into communities. (D) For each exposure a feature vector is created. The gene expression data are filtered to only include genes contained in the prior network (1). The genes are sorted by their up/down regulation and the top (up-regulated) and bottom (down-regulated) 100 genes are selected (2). These 200 genes are mapped onto the prior network partitions (communities). For each exposure a feature vector is created, whose length is equal to the number of detected communities and its values indicate the fraction of the most affected 200 genes falling into each community (3). (E) The feature vectors are used to cluster the compounds for each biological system. (F) The clusters are compared between the biological systems, via a jaccard index.

the same pathway) or on a higher level, such as that genes are associated with the same or closely related phenotypes. A similar approach is applied in multi-omics, where data from different omics technologies are combined in order to generate a more complete view of the analyzed data (Serra *et al.* 2015, Rappoport and Shamir 2018, Mitra *et al.* 2020). The data used to create the prior network is described in the Supplementary Materials (Methods—Prior Network Data Collection). Which has been integrated into a Knowledge Graph framework (Pavel *et al.* 2022), the Unified Knowledge Space (UKS), which has been previously described in Pavel *et al.* (2021a,b) and Federico *et al.* (2022).

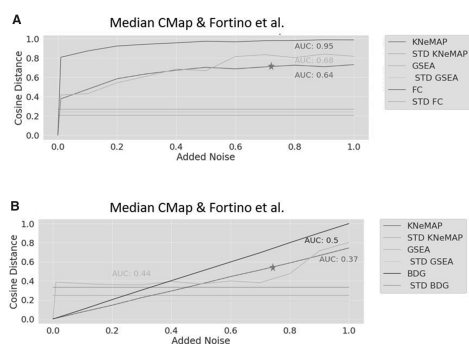
#### 2.1.1.1 Prior network

For each of the data types collected (Paralog, Homolog, Protein Family, Protein Sub-Family, Chemical associations, Disease associations, Pathways, Biological Process, Molecular Function, Cellular Component, PPI) a single gene-gene similarity network was created (Fig. 1B1). For data representing gene-gene edges in the UKS, such as contained in the protein-protein interaction layer a gene-gene similarity network was constructed by retrieving the interactions and assigning as weights the number of data sources supporting this edge. This approach of unifying gene networks has already proven to be effective, as described in Pavel *et al.* (2021a,b). The other type of data, representing gene-entity edges, such as gene-disease associations or gene-pathway associations were converted into a gene-gene similarity network. Here an edge represents two genes that are associated with the same entity (e.g. a disease) and the edge weight represents how many shared entities



the pair of genes has, similarly to the approach described in Federico *et al.* (2022). After the individual gene similarity networks were created, their edge weights were scaled to be in (0,1), where a value close to 1 represents a strong similarity and a value close to 0 represents a weak similarity. This was performed in order to merge the individual networks into a combined gene similarity network. The individual networks were merged in a hierarchical, data-driven fashion. First the individual networks edge similarity was assessed, based on a combined distance on their binary edges. The aim was to first merge data layers, which span similar areas, therefore it was only considered if an edge is present or not and not their computed edge weights, which are first considered in the merging process. The combined distance was created by summing the jaccard distance matrix, the SMC (Simple Matching Coefficient, also known as Rand similarity) distance matrix and a distance matrix computed from the percentage of shared edges (1-fraction of shared edges) (Pavel *et al.* 2021a,b). All three distance matrices were weighted equally and the resulting distance matrix was scaled to be in (0,1). On this combined matrix, hierarchical clustering was performed with `scipy.cluster.hierarchy.linkage(method="ward")` (Virtanen *et al.* 2020), resulting in three main clusters as shown in Supplementary Fig. S1. The networks in the individual clusters were merged first, in such a way that their individual edge weights were scaled to all have the same median value and then were added up, a similar approach has been applied in Federico *et al.* (2022). After this was performed for all three clusters the process was repeated for the resulting three new gene similarity networks in order to create one single combined gene similarity network (Fig. 1B2), whose values were again scaled to be in (0,1). The final created network consisted of 22 316 nodes and 213 784 257 edges, which corresponds to a network density of 0.86. The prior network is available at 10.5281/zenodo.7334711.

On the so created weighted gene similarity network community detection was performed (Fig. 1C) with `volta.communities.agglomerative(distance_threshold=0.5)`



**Figure 2.** Median cosine distance between KNeMAP, BDG, GSEA, and FC-based vectors with increased levels of added noise to the gene expression values as well as the selected deregulated genes. (A) Shows the median performance across both datasets for increasingly added noise. (B) Shows the median performance across both datasets for increased perturbation noise added to the top 200 selected most deregulated genes. The cosine distance between the vectors was computed from the gene expression data with different noise levels or the set of selected deregulated genes and the baseline (noise = 0). The noise levels are on the x-axis, the mean cosine distance on the y-axis. The stars are indicators of the KNeMAP line, used to improve inclusivity of the figure.

(Pavel *et al.* 2021a,b), which performs agglomerative clustering on the networks adjacency matrix using its edge weights (similarities). In order to identify genes that are highly similar in different data layers but not to generate large groups of genes, we aimed at a community distribution of many small-scale communities. In comparison to other community detection algorithms available in VOLTA (Pavel *et al.* 2021a,b), `volta.communities.agglomerative()` showed a partitioning closest to the desired community distribution. The final network partitioning consisted of 1466 communities with a mean size of 15.2 genes per community. The network partitioning is available at <https://github.com/fhaive/KNeMAP/tree/main/data>.

## 2.2 Feature vector creation

The MOA of a compound can be defined as the list of most deregulated genes (Federico *et al.* 2020, Serra *et al.* 2022b). Thus, KNeMAP compares the drug induced transcriptomic alterations by means of a feature vector, capturing the similarity (gene groups on the prior network) between the most deregulated genes. Additionally, in previous analysis of the CMap dataset, it has been suggested that a subset of affected genes is enough to describe the data instance (biological system + exposure) (Struckmann *et al.* 2021). For each data instance, the genes are sorted by their FCP ( $\pm \log FC * -\log(Pval)$ ) score. The top 100 most positive deregulated genes and the top 100 most negative deregulated genes (Fig. 1D2), which are represented in the created gene similarity network (Fig. 1D1), were selected. Supplementary Figure S4 outlines the correlation and distance between feature vectors for different gene set sizes in combination with the variability of these values. The selected genes were mapped onto the computed communities of the prior network and for each community the fraction of the 200 genes falling into that community were estimated. Based on these fractions, a feature vector for each data instance was generated, where each bit position describes a community and its value indicates the distribution of most deregulated genes across them (Fig. 1D3). The script to compute the vectors is available at <https://github.com/fhaive/KNeMAP>.

## 2.3 Similarity of the exposures based on the deregulated genes in a binary feature vector

To compare the KNeMAP method, to a commonly used gene-based method (Scala *et al.* 2018, Kinaret *et al.* 2020a, Saarimäki *et al.* 2020, Kinaret *et al.* 2021, Serra *et al.* 2022a), a binary gene vector (BDG) for each data instance was created. To create this vector, the same 200 genes for each instance, as used in the KNeMAP feature vector, were selected. In a gene wide vector (11 868 genes were measured) a value of 1 was set if the corresponding gene at this position is in the set of 200 most deregulated genes of that specific data instance, else a value of 0 was set.

## 2.4 Similarity of the exposures based on the FCP values in a FCP feature vector

We also compared KNeMAP to a vector making use of all gene FCP values of all common measured genes. For each compound exposure on each system, the gene FCP values were collected into a feature vector (FC). A clustermap (Supplementary Fig. S2A), indicating similarities between sample pairs was computed with `seaborns (Waskom et al. 2018) clustermap(method="ward," metric="euclidean")`. In

addition, the Pearson correlation between all pairwise samples of two biological systems were computed and are displayed in Supplementary Fig. S2B. These two plots show the correlation between instances based on the gene expression fold changes.

## 2.5 Similarity of the exposures based on the GSEA values in a GSEA feature vector

As a third comparison we selected a GSEA (Subramanian *et al.* 2005)-based comparison for KNeMAP, as a more complex and computationally expensive methodology. This approach is in accordance with the method selected by Iorio *et al.* (2010), who used this metric to compute distances between compounds on the CMap dataset. Since KNeMAP, FCP, and BDG are all vectors to describe the alteration profile of a compound on a specific biological system, we computed a GSEA-based vector to describe a compound exposure. For each compound the same top 200 most deregulated genes were selected and used in a GSEA to map against the ranked gene lists (by their FCP) of all the other compounds in a biological system. The GSEA was computed with the blitzGSEA python package (Lachmann *et al.* 2022). The enrichment *P*-values were used to create a feature vector that describes the enrichment of a compound with respect to all other compounds exposed on the same biological system.

## 2.6 Method comparison

### 2.6.1 Comparing compound similarities to prior knowledge

To evaluate KNeMAP's performance to other methods, we compared the numerical correlations and similarities based on their distributions as well as with respect to both functional and structural prior knowledge. In addition, we investigated how susceptible to added noise the four methods are. A comparison between KNeMAP, the BDG vectors, the GSEA vectors as well as the FC vectors was performed. The pairwise Pearson correlations and Cosine distances on all three biological systems were computed and their distributions set side by side.

In addition, we compared the four methods based on their ability to identify functional similar compounds. Since the biological system can have a strong impact on the gene expression profiles (Mullard 2018), we focused on identifying similarities on the same biological system rather than between them in order to minimize system dependent biases towards our validation. Our method validation is based on the assumption that drugs with a similar effect should be more similar in their feature vectors than other drugs on the same biological system. In order to describe compound similarity we retrieved ATC (Anatomical Therapeutic Chemical) codes, where possible for compounds in the CMap dataset. ATC codes are unique identifiers assigned to a drug, which is based on the organ it affects as well as how it works. Where the first level describes its anatomical group, the second a drug's therapeutic group, the third level its pharmacological group, the fourth a drug's chemical group and the last level its chemical substance ([https://www.whocc.no/atc\\_ddd\\_index/](https://www.whocc.no/atc_ddd_index/)). For 312 drugs respective ATC codes could be retrieved (Supplementary File S1). We used the Pearson correlation to compare the two vectors, as suggested by (Struckmann *et al.* 2021), where it was shown to be the highest performing metrics (out of 26) on the L1000 datasets (CMap 2) (Subramanian *et al.* 2017) in identifying the same chemical across different exposures, which vary in system exposed on

or dosage used. To adjust the method to our data, where we only have one exposure of a compound for each biological system, we used the ATC classes to group compounds together. For each compound, *c*, the other compounds were ranked by their similarity to *c*, based on the four different feature vectors (KNeMAP, BDG, GSEA, and FC), and the top *x* (ranging from 1 to the number of compounds for which an ATC code could be retrieved) compounds were selected. Then it was counted how often compounds with the same ATC Code Class (Level 3) were in the top *x*. This value was divided by the total number of the ATC class in the dataset, in order to limit biases through more represented ATC classes. For each *x* all values for each *c* were summed up and displayed in Supplementary Fig. S5. To compare how the methods performed, not only for a specific biological system, we computed the mean for each method across all three biological systems. This allows us to evaluate which method shows the "best" performance on average. The average performance is displayed in Supplementary Fig. S5D. The best performing method is determined by comparing their area under the curve (AUC) scores, where the highest AUC score indicates the best performance. For the NANOSOLUTION data the same metric was performed, however instead of using ATC codes, the core material as well as the ENM shape were used as shown in Supplementary Fig. S9.

In addition, we computed the similarity (based on KNeMAP, the BDG vectors, the GSEA vectors, and the FC vectors) between each compound pair, ranked these pairs based on their similarities and compared the rankings to a similarity computed from the chemical structures. We retrieved the (canonical) SMILES for all CMap compounds, where available, from PubChem (Kim *et al.* 2019, Sayers *et al.* 2022). For each compound pair, in the CMap dataset, (for 450 compounds SMILES were available) the Levenshtein distance, which is the minimum number of character edits needed to make two strings identical (Miller *et al.* 2009), was calculated and the compound pairs were ranked accordingly. This ranking was used as a reference ranking to which the KNeMAP, BDG, GSEA, and FC-based similarity rankings are compared to. Between KNeMAP, BDG, GSEA, and FC, we computed the cosine distance for all compound pairs. Only compounds that had an associated SMILES were considered. These pairs were ranked on their cosine distance. For each method we selected the top *x* (2–200) compound pairs and computed the rank difference between its rank and the SMILES-based rank. The mean of these values was computed and the results are plotted in Supplementary Fig. S6, the curves are compared by means of their AUC of which a lower value indicates more agreement with the SMILE-based ranking. In addition, we computed the jaccard index based on the top *x* (1–1000) pairs and compared the performance of all four methods via their AUC scores, of which a high AUC indicates an overall higher jaccard index (Supplementary Fig. S6). This allows us to evaluate the compound pair similarities against a biological system and exposure indifferent factor, the compound structure. To compare how the methods fare not only for a specific biological system, we computed the mean for each method across all three biological systems. This allows us to evaluate which method shows the "best" performance on average. The average performance is displayed in Supplementary Figs S6D and S7D. In addition, we also computed the rank difference for each method's top 20 compound pairs with the SMILES-based ranking. The density

plots of these values, for each biological system, are displayed in Supplementary Fig. S8. For the Fortino *et al.* data, instead of SMILES, functional descriptors of the ENMs, as downloaded from ([https://github.com/fhaive/metanalysis\\_toxicogenic\\_data/](https://github.com/fhaive/metanalysis_toxicogenic_data/)) were used. Only descriptors available for all ENMs were considered and the cosine distance was estimated between each ENMs descriptor vector of which their pairwise ranks were used the same way as the chemical SMILE-based ranks.

### 2.6.2 Comparing the impact of added noise between the methods

To investigate how the three different methods are reacting to added noise to the data, two different experiments were performed. First different variations of noise were directly added to the batch corrected gene expression data, from which the  $\pm \log_{FC} * -\log(Pval)$  (FCP) scores, as described previously, were calculated. Noise was added per sample, drawn from a Gaussian distribution with mean = 0 and standard deviation levels of 0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 1. For each noise level the KNeMAP, GSEA, and FC vectors, as described previously, were calculated. For each compound, its cosine distance between the added noise levels and the baseline (no noise added to the gene expression vector) was estimated. The mean cosine distance for each noise level across all compounds of a biological system were calculated, together with the average standard deviation (change) across the noise levels, which provides an indication on how much the cosine distance is affected by increasing noise. The cosine distance instead of the Pearson correlation was selected, since we wanted to measure the effect (distance) the different noise levels have with respect to the baseline (noise = 0). For the second experiment the selected 200 most deregulated genes were permuted. Each gene in the selected 200 genes, with a probability of 0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 1 was replaced by another random selected gene from the whole list of measured genes. From this the KNeMAP, GSEA, and BDG vectors were estimated and the cosine distance to their baseline vectors (noise 0) calculated as described in the previous experiment. The results are displayed in Fig. 2, Supplementary Figs S13 and S17.

### 2.7 Stability of KNeMAP vector across different biological systems

To investigate the stability of KNeMAP with respect to differences in steady state gene expressions between different biological systems, we compared the KNeMAP fingerprints computed on different sets of genes. Exposures on different biological systems are known to be different, which partially is caused by the differences in steady state gene expression. To showcase that KNeMAP is robust to such change, we compute the KNeMAP fingerprints based only on the genes that are not differentially expressed as well as only differential expressed genes between the control samples of the individual CMap cell systems. A gene was differentially expressed, if it was classified as differentially expressed between at least one cell line pair. Differential expression analysis was performed with `limma()` (Ritchie *et al.* 2015), as already described in the Supplementary Materials for the pre-processing of the CMap dataset. We then computed the cosine distance between each compound pair on a biological system for both types of vectors and then estimated the difference in cosine distance for each compound pair. The distribution of differences is plotted

in Supplementary Fig. S21, showing that there is a minimal change in pairwise distance between the fingerprints computed based on the complete gene vector or only when taking stable genes between all biological systems into account, due to the independence and multi-dimension of the prior gene-gene network.

### 2.8 Individual analysis of the CMap and Fortino *et al.* dataset

Transcriptomics profiles alterations induced by compound exposure under different experimental conditions (e.g. biological systems, exposure time) can vary strongly (Kinaret *et al.* 2017, Fortino *et al.* 2022). In addition data biases can be present, e.g. due to technical differences, batch effects or to underlying differences in the biological systems (Supplementary Fig. S3) (Federico *et al.* 2020, Serra *et al.* 2020). Therefore we decided to analyze, for the CMap dataset, the three different biological systems independently from each other and merge their results in order to identify similarities between the systems. Analyzing the biological systems independently, allows us to compare the MOA of the exposures detached from the underlying data and in result minimizes data and system related biases, which has been suggested to be an issue of the CMap dataset (Lim and Pavlidis 2021). We performed the same analysis pipeline for the two biological systems available in the Fortino *et al.* dataset. The analysis methodology is described in detail in the Supplementary Materials (Methods—Comparison of the Biological Systems).

### 2.9 Comparative analysis between the CMAP and Fortino *et al.* dataset

To showcase the capability of KNeMAP to compare transcriptomic alteration profiles across datasets, we performed a comparative analysis between the transcriptomic profiles induced by the ENM and drug exposures. Thus, for each ENM in the Fortino *et al.* data we retrieved the most similar drug in the CMap dataset. For each exposure instance between the Fortino *et al.* data and the CMap data we computed the cosine distance between their KNeMAP feature vectors, then ranked the drugs according to their similarity to a nanomaterial exposure. For each nanomaterial–drug pair the mean rank was estimated and the highest ranked drug was selected for the ENM. It is important to note, that when the same prior network is used, KNeMAP offers the possibility to compare different datasets without the need to recompute or adjust the computed feature vectors.

## 3 Results

We developed KNeMAP, a novel methodology for comparison of transcriptomic profiles. We showcased the effectiveness of our method by analyzing the Connectivity Map (CMap) dataset (Lamb *et al.* 2006) and the Fortino *et al.* (2022) dataset. The CMap dataset is a popular reference database for drug-induced expression profiles and combines chemical exposures over three cell lines of which 11 868 genes are measured across all three biological systems. The diversity of CMap makes it a suitable dataset for the identification of groups of chemicals that act similarly on different biological systems, which are challenging to identify with traditional gene-based methods. In Supplementary Fig. S3 the steady state gene expression profiles of the three different cell lines are outlined, which are very different. The Fortino *et al.*

dataset comprises transcriptomic profiles of different nano-materials exposed on two human cell lines (THP-1 and BEAS-2B). The materials vary in core material as well as in their surface chemistry. We evaluated KNeMAP against three existing methods: BDG, GSEA, and FC, by comparing the similarity of transcriptomic profiles calculated with the three methods against similarities computed with independent data layers such as the chemical structure and functional knowledge.

### 3.1 KNeMAP-based similarities better resemble those computed from prior knowledge

To evaluate the performance of KNeMAP, we investigated how it performs with respect to prior knowledge. Since prior knowledge was not equally available for all compounds, these metrics were only computed for compounds where the considered prior knowledge was available. To evaluate the method's capability in identifying structurally similar CMap compounds, pairwise compound similarities were estimated and their rankings compared to compound pair rankings based on KNeMAP, BDG, GSEA, and FC-based vectors. Supplementary Figures S6 and S7 showcase the improvement in agreement to the structural-based ranking for KNeMAP. While differences in performance between the biological systems could be observed. On average (Supplementary Figs S6D and 7D) KNeMAP is in more agreement with the structural-based ranking, which is indicated by lower AUC values (the difference to a structural-based ranking is measured) in Supplementary Fig. S6, a higher AUC values in Supplementary Fig. S6 and a shift of the distribution to the left in Supplementary Fig. S8.

Supplementary Figure S5, showcases the performance of KNeMAP in comparison to BDG, GSEA, and FC in identifying functionally similar CMap compounds. Functional similarity of compounds was determined based on their ATC level 3 codes. However, on average the performance across all three systems is very similar between the methods. For the Fortino *et al.* data, KNeMAP outperforms the other methods on average on identifying ENMs with the same shape (Supplementary Fig. S9E), while GSEA and FC show stronger performance in identifying ENMs based on their core-material (Supplementary Fig. S9F). This suggests that it is advisable to select a metric based on the task to be performed and data quality available. While for the molecular descriptor-based ranking KNeMAP was outperformed by FC for the difference in rankings and BDG for the jaccard index, it performed second best for both methods, overall showing the most stable performance, as displayed in Supplementary Fig. S10.

### 3.2 KNeMAP reduces the noise associated to transcriptomic studies and improves the retrieval of similarity patterns

To show the improvement on the overall comparability of the investigated datasets and to investigate the impact KNeMAP has on the overall similarity distributions, we compared the within dataset distance and correlation by means of the Pearson correlation and cosine distance.

When comparing the Pearson correlation and cosine distance distribution values for each compound pair on each biological system (Supplementary Figs S11 and S12) for the FC vectors, the BDG vectors, the GSEA vectors and KNeMAP, it can be observed that while the BDG and FC-based values show a similar narrow peaked distribution at 0 and 1

respectively, KNeMAP and GSEA yield a broader distribution shifted to the right and left respectively, while GSEA shows a strong difference in shape between the data-sets in contrast to the other three methods. This indicates a shift in similarity/correlation between the exposures, which is not observable based on traditional methods, making this previously difficult dataset easier to analyze and to identify similarities between exposures by reducing the noisy peak observable with the other two methods.

As shown in Fig. 2, KNeMAP is less impacted on average by increasingly added noise to the gene expression values in comparison to the FC and GSEA-based cosine distance. The same applies to KNeMAP in comparison to BDG and GSEA when impacting the selected deregulated genes, which is indicated by its overall lower AUC score.

In Supplementary Figs S14, S15, S18, and S19 the plots are shown for selected compounds, Supplementary Figs S13 and S17 show the performance for each individual biological system as well as the median for each dataset and Supplementary Figs S16 and S20 showcases the standard deviation distribution for each biological system for the cosine distance against its baseline (noise = 0). Next to the overall better AUC scores that KNeMAP achieves (Fig. 2), it can be observed that KNeMAP, FC, and BDG are relatively stable across all five biological systems with respect to their AUC scores, while the performance of GSEA varies strongly across biological systems (Supplementary Figs S13 and S17).

### 3.3 Comparison of transcriptomic profiles across different cell lines identifies compounds with a system dependent similar mechanism of action

Through the clustering of the compounds (Fig. 1E) across the three different biological systems of the CMap dataset, based on KNeMAP, we were able to identify a set of 38 drugs (Supplementary Fig. S22) that behave similarly when exposed on the same biological system (Fig. 1F). From now on, we consider these 38 chemicals during further analysis. Given the low correlation between the individual MOAs (Supplementary Fig. S2A), we hypothesize that these drugs might have different responses in different systems, while showing similarities when exposed to the same cancer cell lines. It is often observed that molecular heterogeneity across cancer cell lines causes differences in response to the same drug, possibly offering a biological explanation to the observed phenomenon (Dagogo-Jack and Shaw 2018). When clustering the individual treatments, it is apparent how they group by the exposed biological system (Supplementary Fig. S23), rather than by drug. Therefore, we investigated possible characteristics of the 38 drugs that would be responsible for their similar behavior. When addressing their therapeutic indications, 33% were antimicrobial drugs, 15% cardiac glycosides (antiarrhythmic agent), 10% hsp90 inhibitors, and 10% antipsychotic (Supplementary Fig. S26). Although all these drug classes have been already repurposed for various cancer treatments, no specific primary molecular target or pathway could justify their similar activity. Therefore, we hypothesized that the chemical structure may be responsible for the observed phenomenon. Through scaffold analysis (Supplementary Table S2) we were able to identify high level scaffolds statistically enriched in this set of drugs (Supplementary Fig. S27) that can interact with membranes, cytoskeleton and alter the redox state. All these targets are very sensible in cancer cell lines, and when targeted they

ultimately induce a cytostatic or cytotoxic effect. We further explored the structure information to identify other compounds that may show the same or similar behavior when exposed on the same biological systems (Supplementary Table S3).

To showcase the functionality of KNeMAP, we also applied this approach to a set of ENMs exposed to two different cell lines. As in the first case study, our approach was able to highlight a cluster of hazardous nanoparticles (gold and quantum dots with various functionalizations) with peculiar optical and electronic properties (Supplementary Fig. S28). It is known that physicochemical characteristics of nanomaterials affect the induced biological response, possibly explaining the observed similarities across cell lines (Liu *et al.* 2006, Ellis *et al.* 2020). A detailed description of the analysis results and the identified drugs can be found in the Supplementary Materials (Results—Comparison of Transcriptomic Profiles Across Different Cell Lines Identifies Drugs with a System Dependent Similar Mechanism of Action and Description of the Identified Nanomaterials).

### 3.4 Identifying drugs and nanomaterials with a similar mechanism of action

Through the comparison of the KNeMAP fingerprints of the Fortino *et al.* data with the CMap data, we identified for each nanomaterial the chemical compound with the most similar MOA across all biological systems. All identified pairs are listed in Supplementary Table S5 and detailed descriptions of selected pairs are provided in the Supplementary Materials (Results—Identifying the Most Similar Chemical for each Nanomaterial Based on their Mechanism of Action). For example a copper oxide nanomaterial was found to act similar to Lycorine and both have been shown to affect acetylcholinesterase and in result the nervous system (Sezer Tuncsoy *et al.* 2019, Kola *et al.* 2023).

## 4 Discussion/conclusions

We propose KNeMAP as a new knowledge-driven method to compare transcriptomic profiles. In comparison to other methods, which focus on individual genes, KNeMAP groups genes into a “similarity group,” which allows to compare expression profiles in a higher-level manner than when comparing genes individually. We showed that a network mapping-based approach is able to identify similar compounds in higher agreement with functional as well as structural prior knowledge, when compared to the BDG, GSEA, and FC methods. In addition, it is able to reduce the observable noise in the data, which makes the dataset easier to analyze and allows it to identify patterns. KNeMAP can be especially suitable for datasets where data from different systems and with different exposure parameters are compared. In this work, the KNeMAP was applied on the CMap (Lamb *et al.* 2006) dataset as well as the Fortino *et al.* dataset (Gallud *et al.* 2020, Kinaret *et al.* 2021) and we were able to identify a set of compounds that always show a similar response between each other on the same biological system, even though their response may vary across biological systems. While the identified CMap compounds have different therapeutic uses and molecular targets they all have been linked to similar effects on cancer, and have often been repurposed for oncological treatments. Since they do not share most of the molecular mechanism, a more traditional

comparison between differentially expressed genes would have not identified this commonality. The underlying differences of the biological systems can explain the differences in expression patterns between the biological systems for similar compounds, suggesting that these compounds affect the cancer cells differently but always in a similar manner between each other (on the same biological system). In order to make statements about the comparability or the behavior of these compounds on non-cancer related biological systems, further analysis needs to be done, showcasing again how important it is to understand the comparability between biological systems with respect to chemical safety assessment. Moreover, when compared with three different gene focused approaches, KNeMAP is able to identify similarities between compounds with higher agreement to functional as well as structural information. When comparing transcriptomic experiments, one limitation is given by the fact that the same molecules (e.g. genes) need to be profiled. However, different experiments are often performed on different platforms, with only partially overlapping probes/genes. The KNeMAP approach can be further exploited in this case and be used to compare the datasets since thanks to the fact that genes can be grouped into communities, no one-to-one mapping between the genes is required. We showcase this by comparing the CMap dataset with the Fortino *et al.* dataset by identifying for each nanomaterial the drug with the most similar MOA across all biological systems. Furthermore, KNeMAP is highly flexible with respect to what prior data is used to construct the network, so can, e.g. only a single data layer (e.g. pathways, GO) be used or a subset of layers, as well as to the size of gene communities to be detected (based on the algorithm chosen). This allows a “stricter” or “looser” view on gene similarity as needed based on the data or study. In conclusion KNeMAP is a generic approach, that can be customized with respect to prior information and gene clusters used, to compare noisy transcriptomic datasets.

## Acknowledgements

The authors thank Pia A. S. Kinaret for the discussion of the initial results.

## Supplementary data

Supplementary data is available at *Bioinformatics* online.

## Conflict of interest

None declared.

## Funding

This work was supported by the Academy of Finland [322761]; European Research Council (ERC) programme, Consolidator project “ARCHIMEDES” [101043848]; and the Tampere Institute for Advanced Study (to A.S. and A.F.).

## References

Dagogo-Jack I, Shaw AT. Tumour heterogeneity and resistance to cancer therapies. *Nat Rev Clin Oncol* 2018;15:81–94.

- Ellis GA, Dean SN, Walper SA *et al.* Quantum dots and gold nanoparticles as scaffolds for enzymatic enhancement: recent advances and the influence of nanoparticle size. *Catalysts* 2020;10:83.
- Federico A, Serra A, Ha MK *et al.* Transcriptomics in toxicogenomics, part II: preprocessing and differential expression analysis for high quality data. *Nanomaterials (Basel)* 2020;10:903.
- Federico A, Fratello M, Scala G *et al.* Integrated network pharmacology approach for drug combination discovery: a multi-cancer case study. *Cancers (Basel)* 2022;14:2043.
- Fortino V, Kinaret PAS, Fratello M *et al.* Biomarkers of nanomaterials hazard from multi-layer data. *Nat Commun* 2022;13:3798.
- Fratello M, Cattelani L, Federico L *et al.* Unsupervised algorithms for microarray sample stratification. *Methods Mol Biol* 2022;2401:121–46.
- Freytag S, Gagnon-Bartsch J, Speed TP *et al.* Systematic noise degrades gene co-expression signals but can be corrected. *BMC Bioinformatics* 2015;16:309.
- Gallud A, Delaval M, Kinaret P *et al.* Multiparametric profiling of engineered nanomaterials: unmasking the surface coating effect. *Adv Sci (Weinh)* 2020;7:2002221.
- Gao S, Han L, Luo D *et al.* Modeling drug mechanism of action with large scale gene-expression profiles using GPAR, an artificial intelligence platform. *BMC Bioinformatics* 2021;22:17.
- Iorio F, Isacchi A, di Bernardo D *et al.* Identification of small molecules enhancing autophagic function from drug network analysis. *Autophagy* 2010;6:1204–5.
- Kim S, Chen J, Cheng T *et al.* PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res* 2019;47:D1102–9.
- Kinaret P, Marwah V, Fortino V *et al.* Network analysis reveals similar transcriptomic responses to intrinsic properties of carbon nanomaterials in vitro and in vivo. *ACS Nano* 2017;11:3786–96.
- Kinaret PAS, Scala G, Federico A *et al.* Carbon nanomaterials promote M1/M2 macrophage activation. *Small* 2020a;16:e1907609.
- Kinaret PAS, Serra A, Federico A *et al.* Transcriptomics in toxicogenomics, part I: experimental design, technologies, publicly available data, and regulatory aspects. *Nanomaterials (Basel)* 2020b;10:750.
- Kinaret PAS, Nidika J, Ilves M *et al.* Toxicogenomic profiling of 28 nanomaterials in mouse airways. *Adv Sci (Weinh)* 2021;8:2004588.
- Kola A *et al.* A comparative study between lycorine and galantamine abilities to interact with AMYLOID  $\beta$  and reduce in vitro neurotoxicity. *Int J Mol Sci* 2023;24:2500.
- Lachmann A, Xie Z, Ma'ayan A *et al.* blitzGSEA: efficient computation of gene set enrichment analysis through gamma distribution approximation. *Bioinformatics* 2022;38:2356–7.
- Lamb J, Crawford ED, Peck D *et al.* The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 2006;313:1929–35.
- Lim N, Pavlidis P. Evaluation of connectivity map shows limited reproducibility in drug repositioning. *Sci Rep* 2021;11:17624.
- Liu N, Prall BS, Klimov VI *et al.* Hybrid gold/silica/nanocrystal-quantum-dot superstructures: synthesis and analysis of semiconductor-metal interactions. *J Am Chem Soc* 2006;128:15362–3.
- Marwah VS, Scala G, Kinaret PAS *et al.* eUTOPIA: solUTion for Omics data Preprocessing and Analysis. *Source Code Biol Med* 2019;14:1.
- Miller FP, Vandome A, McBrewhster J. Levenshtein distance: information theory, computer science, string (computer Science), String metric, Damerau? Levenshtein distance, Spell checker, Hamming distance. Alpha Press, Orlando, 2009.
- Mitra S, Saha S, Hasanuzzaman M *et al.* Multi-view clustering for multi-omics data using unified embedding. *Sci Rep* 2020;10:13654.
- Mullard A. Can you trust your cancer cell lines? *Nat Rev Drug Discov* 2018;17:613.
- Pavel A, Del Giudice G, Federico A *et al.* Integrated network analysis reveals new genes suggesting COVID-19 chronic effects and treatment. *Brief Bioinf* 2021a;22:1430–41.
- Pavel A, Federico A, Del Giudice G *et al.* VOLTA: adVanced mOLecular neTwork Analysis. *Bioinformatics* 2021b;37:4587–8.
- Pavel A, Saarimäki LA, Möbus L *et al.* The potential of a data centred approach & knowledge graph data representation in chemical safety and drug design. *Comput Struct Biotechnol J* 2022;20:4837–49.
- Rappoport N, Shamir R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Res* 2018;46:10546–62.
- Raser JM, O'Shea EK. Noise in gene expression: origins, consequences, and control. *Science* 2005;309:2010–3.
- Ritchie ME, Phipson B, Wu D *et al.* limma powers differential expression analyses for RNA-seq and microarray studies. *Nucleic Acids Res* 2015;43:e47.
- Saarimäki LA, Kinaret PA, Scala G *et al.* Toxicogenomics analysis of dynamic dose-response in macrophages highlights molecular alterations relevant for multi-walled carbon nanotube-induced lung fibrosis. *NanoImpact* 2020;20:100274.
- Sayers EW, Bolton EE, Brister JR *et al.* Database resources of the national center for biotechnology information. *Nucleic Acids Res* 2022;50:D20–6.
- Scala G, Kinaret P, Marwah V *et al.* Multi-omics analysis of ten carbon nanomaterials effects highlights cell type specific patterns of molecular regulation and adaptation. *NanoImpact* 2018;11:99–108.
- Serra A, Fratello M, Fortino V *et al.* MVDA: a multi-view genomic data integration methodology. *BMC Bioinformatics* 2015;16:261.
- Serra A, Coretto P, Fratello M *et al.* Robust and sparse correlation matrix estimation for the analysis of high-dimensional genomics data. *Bioinformatics* 2018;34:625–34.
- Serra A, Fratello M, Cattelani L *et al.* Transcriptomics in toxicogenomics, part III: data modelling for risk assessment. *Nanomaterials (Basel)* 2020a;10:708.
- Serra A, del Giudice G, Kinaret PAS *et al.* Characterization of ENM dynamic dose-dependent MOA in lung with respect to immune cells infiltration. *Nanomaterials (Basel)* 2022b;12:2031.
- Serra A, Saarimäki LA, Pavel A *et al.* Nextcast: a software suite to analyse and model toxicogenomics data. *Comput Struct Biotechnol J* 2022;20:1413–26.
- Sezer Tuncsoy B, Tuncsoy M, Gomes T *et al.* Effects of copper oxide nanoparticles on tissue accumulation and antioxidant enzymes of *Galleria mellonella* L. *Bull Environ Contam Toxicol* 2019;102:341–6.
- Struckmann S, Ernst M, Fischer S *et al.* Scoring functions for drug-effect similarity. *Brief Bioinf* 2021;22:bbaa072.
- Subramanian A, Tamayo P, Mootha VK *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 2005;102:15545–50.
- Subramanian A, Narayan R, Corsello SM *et al.* A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* 2017;171:1437–52.e17.
- The Gene Ontology Consortium. The gene ontology resource: enriching a Gold mine. *Nucleic Acids Res* 2021;49:D325–34.
- Virtanen P, Gommers R, Oliphant TE *et al.*; SciPy 1.0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* 2020;17:261–72.
- Waskom M, Botvinnik O, O'Kane D *et al.* mwaskom/seaborn: v0.9.0 (July 2018). [Computer software]. Zenodo. 2018. 10.5281/zenodo.1313201.



