Pathway connections

Doctoral thesis

PATHWAY CONNECTIONS: CONNECTIVITY OF PATHWAY ELEMENTS AND THEIR DIRECTIONS IN BIOLOGICAL MOLECULAR PATHWAY DIAGRAMS

Ryan A. Miller

2022

Pathway Connections: Connectivity of pathway elements and their directions in biological molecular pathway diagrams

DISSERTATION

to obtain the degree of Doctor at the Maastricht University,
on the authority of the Rector Magnificus,
Prof.dr. Pamela Habibović
in accordance with the decision of the Board of Deans,
to be defended in public on
Tuesday, 6th of December 2022 at 16.00 hours

by

Ryan Alexander Miller

Supervisor(s):
Prof. Dr. Chris T. Evelo

Co-supervisor:
Dr. Egon L. Willighagen

Assessment Committee:
Prof. Dr. Frederik-Jan Van Schooten, (chair)
Dr. Rianne Fijten
Dr. Derek Marren, AstraZeneca, United Kingdom
Prof. Dr. Reinhard Schneider, Université du Luxembourg, Luxembourg

To my family and friends

Contents

1 Introduction

Biology studies how parts of an organism cooperate and act like a system. If parts of this system change, then these will affect other parts of the system (*1*). Cancer is a common example of a complex disease with systemic effects (*2*). Effects like disease make the connections and interactions in biological systems a central theme in biology, a theme that can be applied to all biological entities (*3*). Furthermore, some connections within the system will have their interactions directed in a particular direction. This is a sign of causality within a cell or organism. Understanding how the system's parts are connected and their directed nature is essential in the areas like hereditary and acquired diseases, toxicity, growth and regulation, pharmacology, signalling and many more biological phenomena. How elements of cells are interacting, connected and the directions of these connections help us understand how the elements of a biological system work together. This systemic thinking allows researchers and scientists to look beyond single interactions and instead to look at how the system works together. It allows for a holistic approach that each time takes into account bigger chunks of the system to study how they work together. The connections and the direction of these interactions are not only the central theme of this thesis work but also a concept central to understanding biology itself.

Changes in how the system operates by changing parts alone can justify why it is important to study interactions, but directional information also gives more specific information about how changes influence the system and allows the construction of networks (*4*). Causal relationships, as are suggested by directional information for interactions, also allow for the analysis of the effects of different interventions. This goes to the very idea of causality, in that changes to expression of genes can have a causal impact on downstream processes. Networks of gene products do not need to be represented as pathways describing a biological process, but may be a network constructed of biological interactions. In sources such as the STRING database (*5*) or in NetPath (*6*), users are able to construct networks of interacting proteins or study signalling pathways respectively. This is an interaction network that represents protein-protein interactions which lacks the rest of the biological components of a pathway that influence biological functions. Although pathways are more complex and represent more than individual types of interactions, we can use the approaches developed using the work done for this thesis which focuses on biological pathways for interaction modelling and networks of interactions which are used in bioinformatics as a systematic approach to biology.

To provide access to this knowledge, websites for the pathway or biological compound databases have been used and developed for decades now (*7*). These are useful when researching just a couple compounds or interactions. However, the problem comes from when the need arises to scale up the research to dozens or hundreds of entities and their relationships for analysis. This is common when analyzing omics data (*8*).

However, even the digital pathway drawings cannot easily be used if they cannot be accessed and understood by algorithms or software designed to interpret them. This is when web services come into play in order to retrieve from databases the information that is specifically needed to answer the scientists' specific research question. This comes in the form of Simple Object Access Protocol (SOAP) application programming interfaces (APIs) and Representational state transfer (REST) APIs and have developed over the years. SOAP APIs have been used by the National Center for Biotechnology Information (NCBI), European Bioinformatics Institute (EBI), and others (*9–11*). REST APIs have been deployed by providers such as Ensembl (*12*). Web services provide predefined queries that are available for users to programmatically pull information from the databases and to be selective for the information that they are looking to retrieve. Finally, researchers can interact with these resources' databases via a query language that allows the user to purposely design the retrieval of the information they are seeking rather than using predefined queries. There is still a place for all of these forms of interaction with the biological resources, but it shows the progression of how we interact with our data.

Resources like WikiPathways, Kyoto Encyclopedia of Genes and Genomes (KEGG), and Reactome are used by biologists to represent their knowledge of how elements of an organism work together to enable the organism to work and be successful (*13–16*). A diagram of a pathway is a drawing to represent how elements of the pathway connect and interact with each other. There are two major types of pathways found in these resources, and they may also include some specialized research pathway types, for example adverse outcome pathways found in WikiPathways. The two classes of pathways are signaling pathways and pathways for metabolic processes. Signaling pathways are used to describe the relay of messaging information. The signalling pathways are an important communication mechanism for the system to work properly. Metabolic processes are pathways that represent energy production, molecular digestion, and production of structural components for the organism. The metabolic pathways are essential for energy and growth of the organism. The collection of pathways are essential portions of the overall system and network, but relies on a gene or gene product with a known function in order for people to want to include it in a pathway. For example, WikiPathways in 2018 contains 50% of the unique human coding protein genes compared to Ensembl, 66% of all disease causing genes compared to Online Mendelian Inheritance in Man (OMIM), and 71% of genes believed to be involved in human metabolic processes from GO metabolic processes (*13*, *17*, *18*). Insights in biology are the objective of this thesis which can be gained from examining how the elements are connected to each other and how they are directed from one place to another. We measure and make pathway drawings based upon biological assays. The pathway elements can be drawn in a diagram with all the necessary connections but also need to be represented in ways that the connections are computer readable for bioinformatics analysis (*19*).

Over the years, resources have been developed to address the needs of biologists to create links across resources. Biological databases have been created to fulfill niche understanding of biological processes. They include, but are not limited to, HUGO Gene Nomenclature Committee (*20*) for genes name, Universal Protein Resource (UniProt) for protein sequence information (*21*), the Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank (PDB) for protein structures (*22*), and the Human Metabolome Database (HMDB) for human metabolite information (*23*). Identifiers.org is a resolving system for Uniform Resource Identifiers (URIs) in the scientific community (*24*). Mapping services were also developed to identify IDs from different sources to verify if they are describing the same entity (*25*). These help scientists know that the entities that they put into their pathways, such as proteins or metabolites, are what they intend to include.

Found within pathway diagrams are nodes that represent genes, ribonucleic acids (RNAs), pathways, or metabolites. In general, datanodes in a pathway diagram can be connected with a line, a general interaction, in which case the influence of one node upon another might not be known. A more specific case would be when a line is drawn between datanodes with an arrow head that indicates the direction of influence from one datanode to another. The more specific interactions are also able to be drawn in the pathway diagrams such as

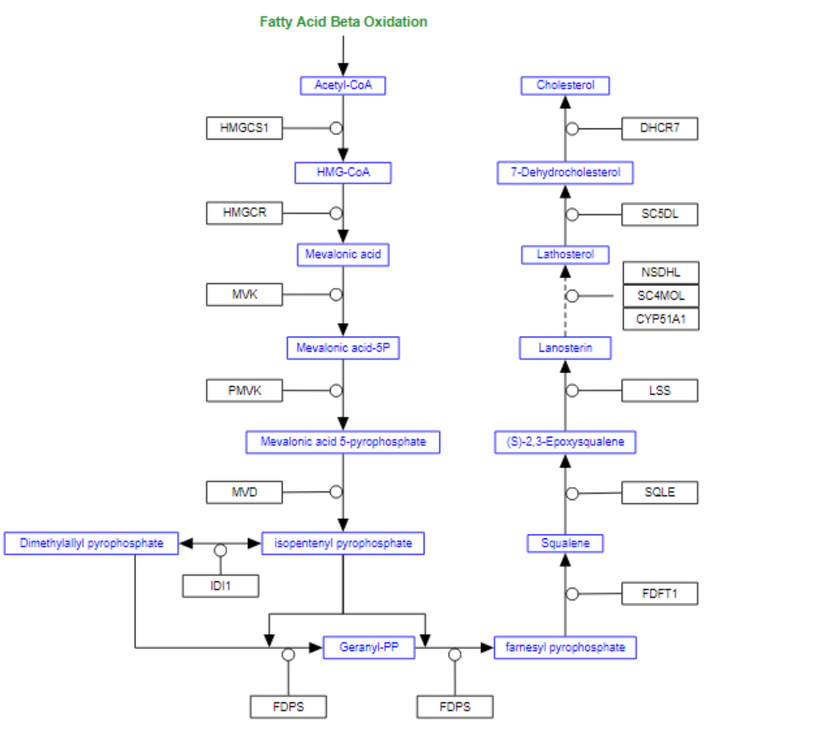


Figure 1.1: Cholesterol Biosynthesis Pathway, wikipathways:WP4141, illustrates how data nodes are represented and how they are connected by edges.

enzymatic conversion of one metabolite form to another with accompanying catalyzing reactions. Inhibition effects can also be shown and are observed in pathways as well as stimulation, and transcription-translation events. Molecular Interaction Map (MIM) and Systems Biology Graphical Notation (SBGN) are two different interaction modeling schemas and common ways of classifying these interaction types (*26*, *27*). Modeling systems like these allow for the semantic capture of interaction information for pathway diagrams.

Knowing how the pathway elements are connected together has a basis in biological experiments and assays to help prove the validity of the interactions for protein-protein interactions as well as for enzyme reactions (*28*, *29*). The connections are one portion of the representation, the other portion being the directional information. One portion of the system has an influence on processes that are found downstream. An example of this would be seen with a drug or small molecule that targets a protein and competitively inhibits its actions and has a direct influence on proteins found downstream of the process that is being targeted. Changes to the system cause changes elsewhere within the system. These elements are complex systems and so changes meant to target one area can have a profound effect on other areas. This can be seen in the case of drugs having unintended side-effects (*30*), which can be severe in nature such as kidney problems, allergic reactions, or decreased immunity. This shows how important it is for a part of the system to work in collaboration with each of the other parts. If the parts do not work, then they can cause larger system wide issues. How the system acts together is an area of study pertinent to understanding biological outcomes.

1.0.1 General Aim

This thesis is centered around the idea of using connectivity data from pathways to further biological understandings and how we can leverage biological pathway connectivity to further the field of biology. Common to all project parts is the theme of the elements of the system being connected to each other and not acting alone. Related to the idea of general connectivity of the system is the idea of directional influences. These influences help the understanding of biological systems, and make it possible to gain new knowledge in biology overall. The pathway diagrams are the main resource used to understand these connections.

These diagrams should allow us to leverage interaction data from various resources to obtain a better understanding of biology. It is applying this principle, of gathering, evaluating and combining data, that marks the principles of this thesis work. WikiPathways is the main resource used, but the biological knowledge comes from many places. WikiPathways allows the user to cite primary literature to a specific data node entity or the edge interactions between them. The November 2016 release of WikiPathways had 22889 citations. The identifier mapping data used by WikiPathways comes from resources like Entrez Gene (*31*), ChEBI (*32*), Ensembl (*33*), HMDB (*23*), and ChEMBL (*34*) as well as others, while the pathway diagrams in WikiPathways can be directly based on literature or interaction databases but they often come from or are inspired by KEGG (*14*), Reactome (*35*), and Pathway Commons (*36*). It is this leverage of information from many sources that makes this thesis possible. Biological information comes from several places and the integration of these sources to make a coherent network aids the study and future studies. This integration of data can be seen in knowledge graphs like Wikidata and applied to the life sciences (*37*, *38*). The integration of data from multiple sources makes other biological questions answerable and to also understand how the system works together.

Unfortunately, it is not always easy to link up all these data sources. But if done properly, it is possible that we cannot only use data from a resource like WikiPathways but are able to combine it with outside sources from the NCBI or EMBL-EBI using appropriate technologies. Linking up the resources to recognize that an entity is the same in WikiPathways and NCBI is only part of the problem. The next problem is how do elements of a pathway link up with these outside sources and what sort of inferences can we gain from these connections. While identifier mapping addresses much of the interoperability for the biological data nodes, for interactions this is not fully explored yet. How to describe and use these interactions then ends up being the bulk of the work done here.

1.0.2 Outline of the Thesis

If we want to use WikiPathways content to analyze a full system in an automated or semi-automated way, we need to make information from WikiPathways into a more interoperable format. The semantic web formalizes interoperability and is explored as a solution for pathways in **Chapter 2**. This chapter explores the modeling of WikiPathways in the Resource Data Framework (RDF), allowing us to explore the content of WikiPathways. Monthly updates to this machine readable information in RDF are provided on the WikiPathways SPARQL endpoint (http://sparql.wikipathways.org/). Interactions are challenging here: graphical lines from a diagram need to be accurately and consistently turned into a form that is able to be queried and stored. The RDF representation of the WikiPathways dataset is a fundamental step toward making the data from WikiPathways available to the scientific community in a form that fits many research interests. These updates to the RDF in the SPARQL endpoint, allow for continuous use of relevant queries for user research that benefits from the latest pathway and pathway model updates.

One of the benefits of turning the WikiPathways graphical information into semantic information is that it allows for WikiPathways data to be integrated with data from other sources. Being able to query information from WikiPathways and merge it with data retrieved from, for instance, ChEMBL or ChEBI allows for taking advantage of this information to answer new questions relevant to biology. This can for instance be used to combine information about how the gene products are interacting and how compounds associated with one gene product may influence that interaction. This has implications in toxicology and pharmacology. The same kind of combined approaches can be useful to study disease development and diseases where there is no clear disease - gene associations. For purposes like this, WikiPathways can be connected to resources like DisGeNet (*39*). Being able to integrate WikiPathways data with other biological and chemical sources from the internet allows for more advanced systems biology studies to be performed. It is this ability to integrate WikiPathways with other resources, that I evaluated in this thesis work, that is an important contribution to biological understanding.

With the overall idea of the semantic web version of WikiPathways done, the next feature needed is an accurate way to semantically describe pathway interactions. Therefore, we set out to explore how to explicitly explain how interactions are modelled and represented in WikiPathways, as described in **Chapter 3**. Here we study how the interactions in the graphical diagrams can be modelled and represent the connections between elements of pathways. The results of this work are needed to provide descriptions for the interactions that will be used later in the project. The objective being to explore if the WikiPathways RDF semantic and connectivity data can be used to answer questions in biology. There were two proposed examples that were used to test this idea. MECP2 was used as an example because it is a protein with implications in a rare disease like Rett syndrome. Our second example, sphingolipid metabolism, is an example of a metabolism pathway, with sphingolipids themselves being important structural elements of cells. These are important examples of how the modelled interactions can be used to explore biological questions of interest using the WikiPathways connectivity and direction information.

Explanation of how the interactions and reactions are modelled in WikiPathways using semantic representation is important to relaying the information and work done by biologists on WikiPathways to other researchers in a way that can be easily queried. This step to create, document, and make available interaction information, through the WikiPathways SPARQL endpoint (https://sparql.wikipathways.org/sparql) and vocabularies (https://vocabularies.wikipathways.org/), is what facilitates the mentioned knowledge transfer. The examples of how the semantic information can be used to answer specific biological questions demonstrate how others can also use the resource to answer their own questions relating to how pathway elements influence each other. Now that **Chapters 2 and 3** have shown how the semantic web approach works for WikiPathways, we can focus on the reuse.

Chapter 4 shows a first use case of this reuse and studies how the WikiPathways connectivity and directional information can be integrated into the Open PHACTS drug discovery platform. This allows the project partners and users to be able to use the integrated data from WikiPathways to explore options only available from a pathway resource. The integrated pathway and connection information enables the querying of the included API to answer questions pertaining to pharmacology. Specifically, directional queries were added to the platform and allowed queries for directional identification of upstream or downstream targets from a protein of interest. Because the WikiPathways semantic data has information about how drug targets and metabolites are connected and interact with each other, the data from WikiPathways is useful for exploration of drug repositioning and repurposing. Making the integration of pathway information an important aspect of the project. The addition of the WikiPathways data also adds richness to the platform in aiding computational drug discovery goals.

The addition of directionality of interactions for the Open PHACTS Discovery Platform is a specific example of how the connectivity information from WikiPathways can be used to answer questions in the area of pharmacological research. This addition does facilitate answering questions that are associated with how proteins influence each other. In the case of pharmacology it allows for querying for targets upstream of the protein of interest and having a similar downstream effect. Being able to move up or down from the original target allows for this type of analysis to be performed for any protein target found in WikiPathways.

Having connectivity information from WikiPathways also means that it is possible to construct networks. Being able to construct pathway based networks by combining pathways into a larger network and evaluating which nodes are active based upon a specific omics dataset is significant for researchers interested in evaluating the most relevant active or changed processes in biological studies. This makes this work a powerful tool and approach for identifying active networks from much larger networks. In **Chapter 5** we explore this application for rare diseases, and discover interaction networks in order to study diseases that are often not yet studied at that level. Rare diseases are hard to study because there are fewer patients with these diseases, being able to find and isolate relevant subnetworks in these diseases is harder because they are less studied with smaller datasets and need further context to make informed conclusions. For example, constructing these subnetworks was used to create a network for Rett syndrome. In this case, MECP2 is a protein of interest in Rett syndrome. MECP2 is a protein that can both activate and repress transcription and is required for neuron development (*40*). MECP2 function is lost for most cases of Rett syndrome with varying degrees of phenotypic severity (*41*). The Gene Expression Omnibus (GEO) was used to find an appropriate transcriptomics dataset for the study. A larger network was created from WikiPathways for humans consisting of a network of all human pathways. A subnetwork of active nodes was identified from the human network for Rett syndrome. The subnetwork is a network specific to this disease's gene expression characteristics.

The significance of identifying active nodes of a larger network from a publicly available dataset, means that it is possible to construct and identify active subnetworks for biological study in places like rare diseases. Rare diseases are less studied than their more common disease counterparts. While rare diseases may have several datasets pertaining to them, you often find these datasets are smaller. This means the underlying system and system effects are less studied too. This basic system information is important for scientists interested in translational research.

Another example of the use in pharmaceutical research is defined by the drug synergy DREAM challenge (*42*), which was hosted by AstraZeneca and Sanger as a community competition to address questions in systems biology. The challenge objective was to use drug data to predict potential drug combination synergies. This shows that connectivity and directional information gained from WikiPathways can also be used in areas like cancer research. The idea being to use prior biological knowledge to address cancer pharmacology. The central hypothesis here is that synergetic effects originate from how the targets they hit are biologically related. This is defined by the interactions in the pathways. In **Chapter 6** we explore how the WikiPathways data can be used in pharmacology and cancer research. The specific challenge we chose was to use data provided by the DREAM challenge organizers to predict synergies without the aid of training data, this challenge was the inspiration for this chapter as further work was done after the challenge had been completed.

Our approach is as follows. One model for prediction is the Loewe Additivity Model that says if two drugs share the same target, the drug combination can be mathematically calculated to be synergistic or additive in nature. The first approach used for the calculations was to take the idea of two drugs sharing a common target and apply it to the idea that these two drugs do not have to share a common target but instead need only share a common pathway. The idea being that if two drugs are active and target two different proteins within the same pathway then each drug is targeting a part of the same biological process. This has a disadvantage that potential drug targets in a pathway may not share the same arm or directed path through the pathway and so may not have the same type of effect and may not have much of an influence on one another. So the next approach was to use specific connections between targets that are in the same common directed branch of the pathway to make connections for potential drug targets. That is, of the two potential drug - target combinations one needs to be directly upstream or downstream, by up to four intermediate steps, of the other and thus they share a common path and the downstream protein is being influenced by one that is upstream of it. This means that the set of possible drug-target combinations is more numerous than in the case of just making calculations with two targets sharing the same target, but the approach is also more specific. This approach targets protein combinations that share a common directed path through connected biological interactions of the two targets, and is more specific than the approach that uses targets that only share a pathway. This makes the last approach of using targets found in the same directed paths through a pathway as the preferred method for proposing potential drug-target combinations. We can use previously known knowledge of pathway connectivity to make predictions about potential drug-target combinations. The specific biological meaning of the interactions are essential here.

Using the semantic information from WikiPathways, and more specifically directed interaction information, to identify possible synergistic drug combinations, displays an important example of the ability of integrating pathway data with data from other resources for use in research. It is an important example of how the pathway diagrams can be used for other applications other than pathway enrichment systems biology approaches. It means that it is possible to use the directional data from WikiPathways to make pharmacological predictions about how they will interact with each other and the system.

The chapters of the thesis illustrate the describing and use of interactions, connectivity, directional components, and demonstrate how it provides an opportunity to explore the intricacies of pathway diagrams and how they are represented in a way that is applicable to human biology. **Chapters 2** and **3** chapter explain the semantification of the interaction knowledge. Next, I show the applications of this approach in several fields such as pharmacology as seen in **Chapters 4** and **6**, cancer research as seen in **Chapter 6**, and network biology as seen in **Chapter 5**. **Chapter 7** will discuss the overall results of this thesis and argue that this approach provides a unique opportunity to use pathway information to better explore biology.

References

1. A.-L. Barabási, N. Gulbahce, J. Loscalzo. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*. 2010. **12** (1): 56–68.
2. J. J. Hornberg, F. J. Bruggeman, H. V. Westerhoff, J. Lankelma. Cancer: A Systems Biology disease. *Biosystems*. 2006. **83** (2-3): 81–90.
3. S. Yi *et al.* Functional variomics and network perturbation: connecting genotype to phenotype in cancer. *Nature Reviews Genetics*. 2017. **18** (7): 395–410.
4. S. Triantafillou *et al.* Predicting Causal Relationships from Biological Data: Applying Automated Causal Discovery on Mass Cytometry Data of Human Immune Cells. *Scientific Reports*. 2017. **7**: 12724.
5. D. Szklarczyk *et al.* STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*. 2018. **47** (D1): D607–D613.
6. K. Kandasamy *et al.* NetPath: a public resource of curated signal transduction pathways. *Genome Biology*. 2010. **11** (1): R3.
7. L. D. Stein. Integrating biological databases. *Nature Reviews Genetics*. 2003. **4** (5): 337–345.
8. B. Berger, J. Peng, M. Singh. Computational solutions for omics data. *Nature Reviews Genetics*. 2013. **14** (5): 333–346.
9. E. W. Sayers *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*. 2010. **39** (Database): D38–D51.
10. A. Labarga, F. Valentin, M. Anderson, R. Lopez. Web Services at the European Bioinformatics Institute. *Nucleic Acids Research*. 2007. **35** (Web Server): W6–W11.

11. T. Katayama, M. Nakao, T. Takagi. TogoWS: integrated SOAP and REST APIs for interoperable bioinformatics Web services. *Nucleic Acids Research*. 2010. **38** (Web Server): W706–W711.

12. A. Yates *et al.* The Ensembl REST API: Ensembl Data for Any Language. *Bioinformatics*. 2014. **31** (1): 143–145.

13. D. N. Slenter *et al.* WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Research*. 2017. **46** (D1): D661–D667.

14. M. Kanehisa, Y. Sato, M. Furumichi, K. Morishima, M. Tanabe. New approach for understanding genome variations in KEGG. *Nucleic Acids Research*. 2018. **47** (D1): D590–D595.

15. B. Jassal *et al.* The Reactome pathway knowledgebase. *Nucleic Acids Research*. 2019. **48** (D1): D498–D503.

16. S. Chowdhury, R. R. Sarkar. Comparison of human cell signaling pathway databases—evolution, drawbacks and challenges. *Database*. 2015. **2015**: bau126.

17. M. Kutmon *et al.* WikiPathways: capturing the full diversity of pathway knowledge. *Nucleic Acids Research*. 2016. **44** (D1): D488–D494.

18. M. Martens *et al.* WikiPathways: connecting communities. *Nucleic Acids Research*. 2020. **49** (D1): D613–D621.

19. P. Shannon. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*. 2003. **13** (11): 2498–2504.

20. B. Yates *et al.* Genenames.org: the HGNC and VGNC resources in 2017. *Nucleic Acids Research*. 2016. **45** (D1): D619–D625.

21. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*. 2018. **47** (D1): D506–D515.

22. S. K. Burley *et al.* RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Research*. 2018. **47** (D1): D464–D474.

23. D. S. Wishart *et al.* HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Research*. 2017. **46** (D1): D608–D617.

24. S. M. Wimalaratne *et al.* Uniform resolution of compact identifiers for biomedical data. *Scientific Data*. 2018. **5** (1): 180029.

25. M. P. van Iersel *et al.* The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. *BMC Bioinformatics*. 2010. **11** (1): 5.

26. A. Luna *et al.* A formal MIM specification and tools for the common exchange of MIM diagrams: an XML-Based format, an API, and a validation method. *BMC Bioinformatics*. 2011. **12** (1): 167.

27. N. L. Novère *et al.* The Systems Biology Graphical Notation. *Nature Biotechnology*. 2009. **27** (8): 735–741.

28. C. E. Khamlichi *et al.* Bioluminescence Resonance Energy Transfer as a Method to Study Protein-Protein Interactions: Application to G Protein Coupled Receptor Biology. *Molecules*. 2019. **24** (3): 537.

29. H. Bisswanger. Enzyme assays. *Perspectives in Science*. 2014. **1** (1-6): 41–55.

30. S. Dasari, P. B. Tchounwou. Cisplatin in cancer therapy: Molecular mechanisms of action. *European Journal of Pharmacology*. 2014. **740**: 364–378.

31. D. Maglott, J. Ostell, K. D. Pruitt, T. Tatusova. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research*. 2010. **39** (Database): D52–D57.

32. J. Hastings *et al.* ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Research*. 2015. **44** (D1): D1214–D1219.

33. D. R. Zerbino *et al.* Ensembl 2018. *Nucleic Acids Research*. 2017. **46** (D1): D754–D761.

34. A. Gaulton *et al.* The ChEMBL database in 2017. *Nucleic Acids Research*. 2016. **45** (D1): D945–D954.

35. A. Fabregat *et al.* The Reactome Pathway Knowledgebase. *Nucleic Acids Research*. 2017. **46** (D1): D649–D655.

36. E. G. Cerami *et al.* Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Research*. 2010. **39** (Database): D685–D690.

37. A. Waagmeester *et al.* Wikidata as a knowledge graph for the life sciences. *eLife*. 2020. **9**: e52614.

38. A. Waagmeester *et al.* A protocol for adding knowledge to Wikidata: aligning resources on human coronaviruses. *BMC Biology*. 2021. **19** (1): 12.

39. J. Pinero *et al.* DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database*. 2015. **2015**: bav028–bav028.

40. S. E. Swanberg, R. P. Nagarajan, S. Peddada, D. H. Yasui, J. M. LaSalle. Reciprocal co-regulation of EGR2 and MECP2 is disrupted in Rett syndrome and autism. 2008. **18** (3): 525–534.

41. F. Ehrhart *et al.* A catalogue of 863 Rett-syndrome-causing MECP2 mutations and lessons learned from data integration. *Scientific Data*. 2021. **8** (1): 10.

42. M. P. Menden *et al.* Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. *Nature Communications*. 2019. **10** (1): 2674.

2
Using the Semantic Web for Rapid Integration of WikiPathways with Other Biological Online Data Resources

Adapted from: A. Waagmeester *et al.* Using the Semantic Web for Rapid Integration of WikiPathways with Other Biological Online Data Resources. *PLOS Computational Biology*. 2016. **12** (6): 1–11.

Abstract

The diversity of online resources storing biological data in different formats provides a challenge for bioinformaticians to integrate and analyse their biological data. The semantic web provides a standard to facilitate knowledge integration using statements built as triples describing a relation between two objects. WikiPathways, an online collaborative pathway resource, is now available in the semantic web through a SPARQL endpoint at http://sparql.wikipathways.org. Having biological pathways in the semantic web allows rapid integration with data from other resources that contain information about elements present in pathways using SPARQL queries. In order to convert WikiPathways content into meaningful triples we developed two new vocabularies that capture the graphical representation and the pathway logic, respectively. Each gene, protein, and metabolite in a given pathway is defined with a standard set of identifiers to support linking to several other biological resources in the semantic web. WikiPathways triples were loaded into the Open PHACTS discovery platform and are available through its Web API (https://dev.openphacts.org/docs) to be used in various tools for drug development. We combined various semantic web resources with the newly converted WikiPathways content using a variety of SPARQL query types and third-party resources, such as the Open PHACTS API. The ability to use pathway information to form new links across diverse biological data highlights the utility of integrating WikiPathways in the semantic web.

2.1 Introduction

Pathway analysis and visualisation of data on pathways provide insights into the underlying biology of effects found in genomics, proteomics, and metabolomics experiments (*2*–*5*). WikiPathways is a pathway repository where content is provided by the community at large (*6*, *7*). In a given pathway, elements like genes, proteins, metabolites, and interactions are identified using common accession numbers from reference databases such as Entrez Gene (*8*), Ensembl (*9*), UniProt (*10*), HMDB (*11*), ChemSpider (*12*), PubChem (*13*) and ChEMBL (*14*). Multiple databases can be referenced to annotate an element of the same semantic type, e.g. Ensembl and Entrez Gene to annotate gene information. Even single studies sometimes use different reference databases to annotate experimental findings. It is common for bioinformaticians to spend valuable time dealing with data mapping issues that impede the actual data analysis and interpretation. In WikiPathways we use the open source software framework BridgeDb (*15*), to help resolve different identifiers representing the same (or related) entities. Capturing a semantically correct description of biological entities and their connections across datasets is the broader challenge that we have to address. The semantic web provides an approach to define entities and their relationships. By explicitly defining these entities and relationships the semantic web can provide a network of linked data (*16*). The Resource Description Framework (RDF) consists of two key components: statements and universal identifiers. Each statement is captured as a triple, consisting of a subject, a predicate, and an object. For example, the following triple defines the glucose molecule as being part of the glycolysis pathway:

$$\underbrace{< Glycolysis >}_{subject}\underbrace{< HasMember >}_{predicate}\underbrace{< Glycolysis >}_{object}$$

The notion of a semantic web surfaces as you link across large sets of triples representing a vast number of objects and diverse types of concepts and predicates. The use of uniform identifiers, or URIs (*17*), provides consistency when specifying subjects and objects. identifiers.org (*18*), for example, provides a clearinghouse for a wide variety of URIs for biological entities in the life science domain. WikiPathways provides identifiers for all its pathways and identifiers.org provides the URI scheme to make these resolvable. Standardized URIs for predicates come from efforts such as the Simple Knowledge Organization System (SKOS) (*19*). For example, our example triple above can be expressed in a more universal way as:

$$\underbrace{http://www.identifiers.org/wikipathways/WP534}_{subject}$$
$$\underbrace{http://www.w3.org/2004/skos/coremember}_{predicate}$$
$$\underbrace{http://www.identifiers.org/chebi/CHEBI:4167}_{object}$$

where each element is uniquely and universally resolvable to a defined concept (glycolysis, "has member", and glucose respectively). Of course, the more human readable information can also be explicitly added by describing the labels in RDF. But that information is also available by resolving the URIs.

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX wp: <http://identifiers.org/wikipathways/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX chebi: <http://identifiers.org/chebi/CHEBI:>
wp:WP534 skos:member chebi:4167.
wp:WP534 rdfs:label
    "Glycolysis and Gluconeogenesis (Homo sapiens)"@en.
chebi:4167 rdfs:label "Glucose"@en.
```

In order to contribute pathway knowledge to the semantic web, we have modeled the content of WikiPathways to form triple-based statements. The interactions and reactions curated at WikiPathways are particularly well-suited to enrich the overall connectivity of the semantic web. Pathways offer a meaningful context for relations between biological entities, such as proteins, metabolites and diseases that are otherwise defined in disparate databases. We report on the conversion process and the development of two new vocabularies essential in capturing the semantics behind pathway diagrams. Finally, we evaluate the use of the semantically linked pathway knowledge through specialized queries and third-party resources, showing how to link WikiPathways with disease annotations (from UniProt (*10*) and DisGeNET (*20*)), with gene-expression values (from Gene Express Atlas) and with bioactive chemical compounds known to affect proteins that occur in pathways (e.g. from ChEMBL).

2.2 Results and Discussion

2.2.1 Pathway vocabularies

There are existing standards to model various aspects of pathway knowledge, such as BioPAX (*21*), SBGN (*22*), MIM (*23*), SBML (*24*) and SBO (*25*). BioPAX and SBO are in fact already available in a Semantic Web-compatible language called OWL (*26*). These standards provide valuable building blocks for our “WP” vocabulary that captures the biological meaning of pathways. However, not all of the graphical annotations, spatial information and other subtleties critical for the visual representation, the intuitive understanding and the usability for data visualisation of the curated content at WikiPathways are captured by these standards. Our “GPML” vocabulary directly reflects these features defined in the XML format, GPML, or Graphical Pathway Markup Language. For example, in GPML, all genes, proteins and metabolites are types of data nodes, which are rendered as a rectangular box with properties capturing among others its position, height, width, label, and external reference. For example:

```
<DataNode TextLabel="Glucose" GraphId="dba83" Type="Metabolite">
  <Graphics CenterX="279.0" CenterY="468.0" Width="112.0"
```

```
          Height="20.0" ZOrder="32768">
  <Xref Database="ChEBI" ID="CHEBI:4167" />
</DataNode>
```

In the GPML vocabulary, used for semantic representation of pathway diagrams, the markup elements and values are described as classes and properties, each with their respective URIs.

```
<http://identifiers.org/chebi/CHEBI:4167> rdf:type gpml:DataNode .
<http://identifiers.org/chebi/CHEBI:4167> rdfs:label "Glucose"@en .
<http://identifiers.org/chebi/CHEBI:4167> gpml:graphId "dba83" .
<http://identifiers.org/chebi/CHEBI:4167> gpml:ZOrder 32768 .
```

The GPML vocabulary, in its current form, is mainly instrumental in the representation of the spatial information captured at WikiPathways. However, as we will describe below it can also be used to convert pathway information from other semantic web resources into a format amenable to being rendered and curated at WikiPathways. Explicit mappings to external (graphical) ontologies are not added, however through plugins such as Pathvisio-MIM (*27*) mappings to graphical notations such as MIM or SBGN, are possible. In an analogous way, the WP vocabulary can be used to capture the biological relations from other pathways in such a way that they can be used in resources using this semantic layer of the WikiPathways RDF. We used this approach for example to make the relations from Reactome pathways available in the Open PHACTS discovery platform (*28*) starting from the converted pathways at WikiPathways.

The WP vocabulary, focusing on biological meaning, issues URIs for biological concepts and disregards layout and other rendering details. Using URIs from this vocabulary allows stating that something is a Pathway, or that a DataNode is a chemical compound or gene product. The vocabulary also captures descriptive elements, such as labels, shapes and lines that help annotate and contextualize the pathway reaction details. The RDF generated consist of terms from the vocabularies developed in this context. This is done to be able to reflect the semantics used in the WikiPathways community. However, to allow integration with external pathway resources—which is the primary objective of this project—we need to link to external ontologies. For the subset of concepts in common with prior vocabularies, such as BioPAX, we utilize the SKOS data model to express a range of similarities from skos:exactMatch to skos:closeMatch (*19*, *29*).

2.2.2 Pathway conversion and queries

With these vocabularies in place, the next step is the actual conversion of GPML files into triples using the GPML vocabulary. Then rules are applied to make the biological meaning explicit using the WP vocabulary. For example a directed interaction is captured in GPML as two "DataNodes", a line and an arrowhead. The "DataNodes" have external references as properties. Rules are then applied to state that a line is a Directed Interaction, with a source and a target. Figure 2.1 contains an example of such a rule based reasoning query that issues triples with URIs from the WP vocabulary.

WikiPathways pathways are regularly curated by a team of volunteers that evaluate their usability for analysis and tag the pathways as "curated". WikiPathways contains 1000 pathways in the curated set across over a dozen species that convert to a total of 1.6 million triples. The triples are loaded in a SPARQL endpoint (http://sparql.wikipathways.org), which allows semantic querying of the data with the SPARQL query language (*30*). RDF, including new and updated pathways, is generated and tested regularly and can be delivered upon request. Updates of the RDF that is available for download and in the SPARQL endpoint are triggered by crucial events, such as Reactome or Open PHACTS data releases. This prevents discrepancies in quality control or curation, due to small differences between (frequent) releases. Example SPARQL queries and their plain language translations are given in Figure 2.2. A broad set of $\sim$ 50 queries is available on the help pages of WikiPathways (*31*).

A federated SPARQL query (*18*) enables querying over multiple SPARQL

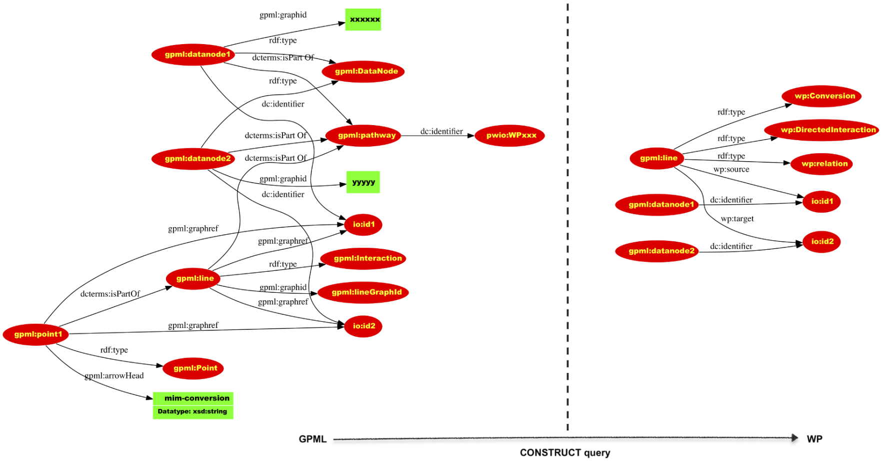
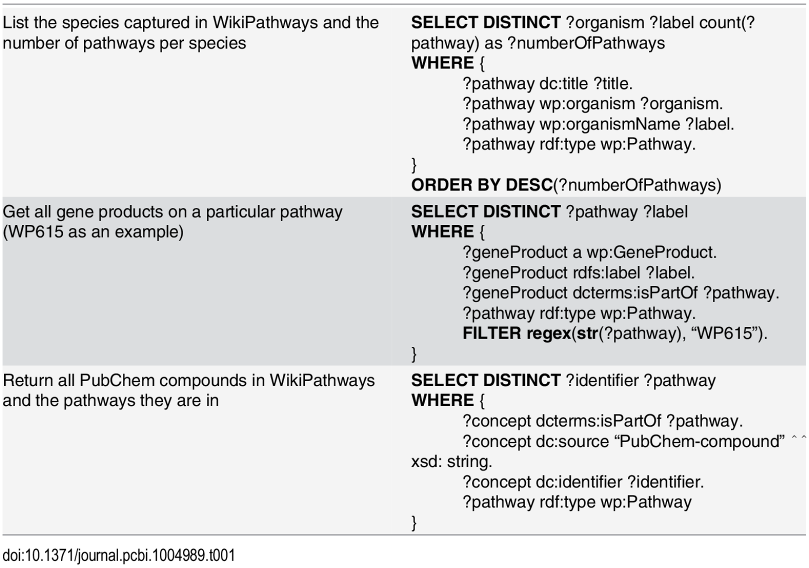


Figure 2.1: A construct query is type of SPARQL query that enables the conversion of one graph pattern to another. Here an interaction described by its spatial properties (GPML) is converted into a semantic representation reflecting its biological interpretation (WP). The SPARQL query is available in the supporting information section.

endpoints. With a variety of SPARQL endpoints available with data on disease annotations (e.g. DisGeNET and UniProt), significantly expressed genes (e.g. EBI Expression Atlas) and drug-target interactions (e.g. ChEMBL), knowledge from these remote SPARQL endpoints can be integrated. Example queries are given in Figure 2.3 and on the help pages of WikiPathways (*32*)

2.2.3 Using linked data in common analysis platforms

Different common analysis platform allow the integration of linked data for future analysis and visualization. One nice example of such a analysis platform is R, a widely used software environment for statistical computing and graphics. R has a SPARQL library (*33*), which enables the import of linked data for further processing in R. This allows running common statistical tests or the creation of different visu-



List the species captured in WikiPathways and the number of pathways per species	**SELECT DISTINCT** ?organism ?label count(?pathway) as ?numberOfPathways
	WHERE {
	?pathway dc:title ?title.
	?pathway wp:organism ?organism.
	?pathway wp:organismName ?label.
	?pathway rdf:type wp:Pathway.
	}
	ORDER BY DESC(?numberOfPathways)

Get all gene products on a particular pathway (WP615 as an example)	**SELECT DISTINCT** ?pathway ?label
	WHERE {
	?geneProduct a wp:GeneProduct.
	?geneProduct rdfs:label ?label.
	?geneProduct dcterms:isPartOf ?pathway.
	?pathway rdf:type wp:Pathway.
	FILTER regex(**str**(?pathway), "WP615").
	}

Return all PubChem compounds in WikiPathways and the pathways they are in	**SELECT DISTINCT** ?identifier ?pathway
	WHERE {
	?concept dcterms:isPartOf ?pathway.
	?concept dc:source "PubChem-compound" ^^ xsd: string.
	?concept dc:identifier ?identifier.
	?pathway rdf:type wp:Pathway
	}

doi:10.1371/journal.pcbi.1004989.t001

Figure 2.2: Example queries handled by the WikiPathways SPARQL endpoint.

alization of linked data. We recently published an R library that interfaces R with PathVisio (*34*) and allows manipulation of pathways and data visualisation on pathways. Figure 2.4 shows up and down regulated genes in Diabetes Mellitus (efo:EFO_0000400, efo:EFO_0001359, and efo:EFO_0001360) in the pathway diagram on insulin signaling in human (*32*). This pathway diagram with color-coding parts indicating up- and down regulated pathway elements, was created by integrating knowledge from two geographically dispersed and independent resources, through a single SPARQL query embedded in a R script, which is available online (*35*).

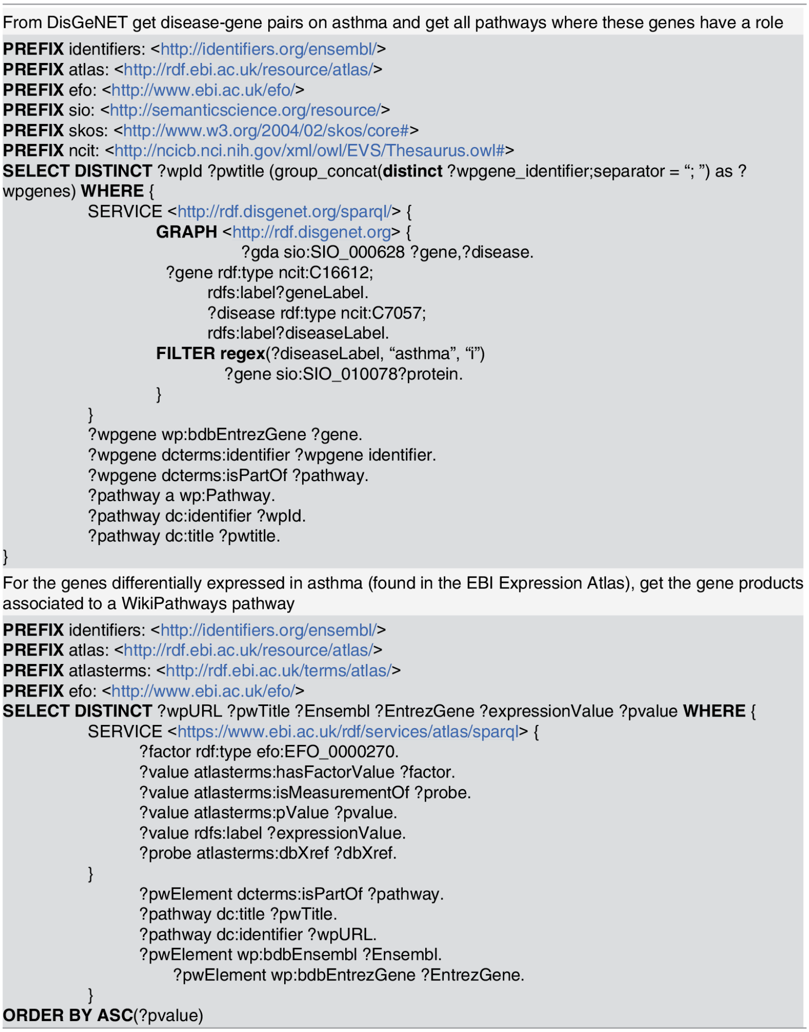
2.2.4 Rosetta stone function

A number of resources provide content from multiple pathway databases, including Pathway Commons (*36*) and NCBIs BioSystems (http:

//ncbi.org/biosystems). While BioPAX in fact is RDF, the NCBI system is not. NCBI BioSystems uses NCBIs native identifiers: GeneId, ProteinId, CID. We thus have a resource with pathways from different origins that are already described in the same way. Since for WikiPathways content we know how the different entities in these resources map to the GPML and WP vocabularies we can now use that to produce RDF using these same ontologies for each of the other pathway resources present in NCBI BioSystems. In fact, we can do the same for Pathway Commons where this approach will lead to an improved version of RDF with explicit mappings to the WP vocabulary. We made a prototype script available on GitHub to be used for this type of conversions from BioSystems (*37*).

2.2.5 Use in discovery platforms

The semantically linked pathway data from WikiPathways RDF have also been integrated into the Open PHACTS discovery platform (*28*, *38*). Open PHACTS delivers and sustains an open pharmacological space using semantic web standards and technologies. The Open PHACTS platform currently provide 51 API methods of which thirteen deliver pathway information (https://dev.openphacts.org/docs). Other information collected in Open PHACTS describes other relationships like drug-target (from ChEMBL) and protein interaction (from UniProt). Having this all in one resource combined with a set of mapping tools allows fast analysis across the domains. By combining Open PHACTS API calls one can, for instance, find all protein targets for a drug and then all pathways that contain these targets.



From DisGeNET get disease-gene pairs on asthma and get all pathways where these genes have a role

```
PREFIX identifiers: <http://identifiers.org/ensembl/>
PREFIX atlas: <http://rdf.ebi.ac.uk/resource/atlas/>
PREFIX efo: <http://www.ebi.ac.uk/efo/>
PREFIX sio: <http://semanticscience.org/resource/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX ncit: <http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#>
SELECT DISTINCT ?wpId ?pwtitle (group_concat(distinct ?wpgene_identifier;separator = “; ”) as ?
wpgenes) WHERE {
          SERVICE <http://rdf.disgenet.org/sparql/> {
                  GRAPH <http://rdf.disgenet.org> {
                            ?gda sio:SIO_000628 ?gene,?disease.
                   ?gene rdf:type ncit:C16612;
                        rdfs:label?geneLabel.
                        ?disease rdf:type ncit:C7057;
                        rdfs:label?diseaseLabel.
                  FILTER regex(?diseaseLabel, “asthma”, “i”)
                          ?gene sio:SIO_010078?protein.
                  }
          }
          ?wpgene wp:bdbEntrezGene ?gene.
          ?wpgene dcterms:identifier ?wpgene identifier.
          ?wpgene dcterms:isPartOf ?pathway.
          ?pathway a wp:Pathway.
          ?pathway dc:identifier ?wpId.
          ?pathway dc:title ?pwtitle.
}
```

For the genes differentially expressed in asthma (found in the EBI Expression Atlas), get the gene products associated to a WikiPathways pathway

```
PREFIX identifiers: <http://identifiers.org/ensembl/>
PREFIX atlas: <http://rdf.ebi.ac.uk/resource/atlas/>
PREFIX atlasterms: <http://rdf.ebi.ac.uk/terms/atlas/>
PREFIX efo: <http://www.ebi.ac.uk/efo/>
SELECT DISTINCT ?wpURL ?pwTitle ?Ensembl ?EntrezGene ?expressionValue ?pvalue WHERE {
          SERVICE <https://www.ebi.ac.uk/rdf/services/atlas/sparql> {
                ?factor rdf:type efo:EFO_0000270.
                ?value atlasterms:hasFactorValue ?factor.
                ?value atlasterms:isMeasurementOf ?probe.
                ?value atlasterms:pValue ?pvalue.
                ?value rdfs:label ?expressionValue.
                ?probe atlasterms:dbXref ?dbXref.
          }
                ?pwElement dcterms:isPartOf ?pathway.
                ?pathway dc:title ?pwTitle.
                ?pathway dc:identifier ?wpURL.
                ?pwElement wp:bdbEnsembl ?Ensembl.
                    ?pwElement wp:bdbEntrezGene ?EntrezGene.
          }
ORDER BY ASC(?pvalue)
```

doi:10.1371/journal.pcbi.1004989.t002

Figure 2.3: Example federated queries handled by the WikiPathways SPARQL endpoint.

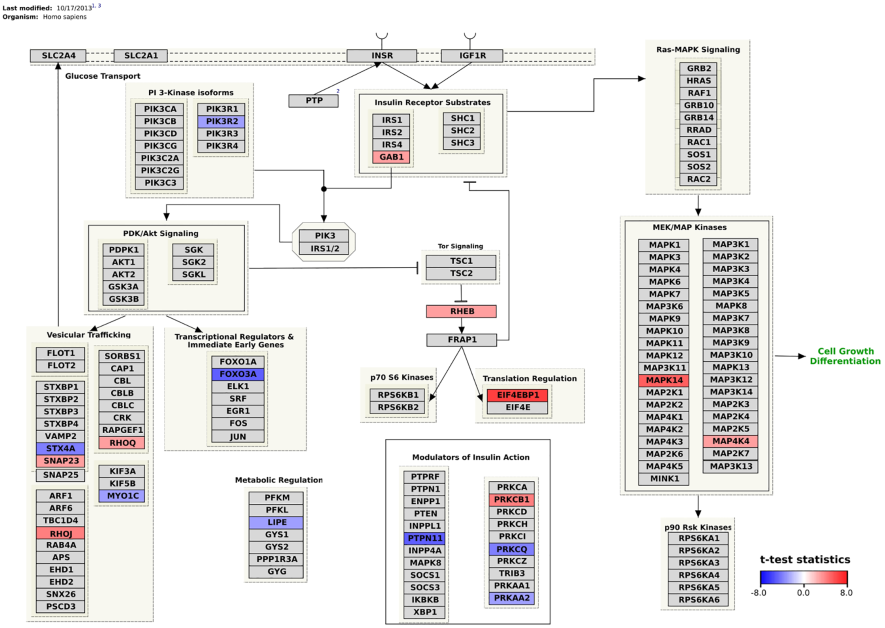


Figure 2.4: The colored boxes represent genes which are up (red) or down (blue) regulated in diabetes mellitus. PIK3R2, MYO1C, PRKAA2, LIPE are down regulated in pre-diabetes. STX4A is down regulated in type 1 diabetes longstanding. PRKCQ, PTPN11, FOXO3A are down regulated in type 2 diabetes. GAB1, RHEB, MAP4K4, SNAP23 are up regulated in pre-diabetes. RHOJ, PRKCB are up regulated in type 1 diabetes recent onset. MAPK14UP, EIF4EBP1 are up regulated in type 1 diabetes clinical onset. From these 17 up or down regulated genes, 9 are being reported as being in the top 10 disease and phenotype associations for the selected gene in DisGeNET (i.e. PIK3R2, PRKAA2, LIPE, STX4A, PRKCQ, FOXO3A, MAP4K4, SNAP23, and PRKCB) (Gene-disease association data were retrieved from the DisGeNET Database, GRIB/IMIM/UPF Integrative Biomedical Informatics Group, Barcelona. (http://www.disgenet.org/). 04, 2016)

2.3 Materials and Methods

Use of Open PHACTS RDF guidelines In collaboration with partners in the Open PHACTS project, we proposed guidelines for presenting data as RDF (*39*), most of that can be considered as general guidelines to produce RDF in the biomedical domain. The guidelines consist of a prerequisite and 11 steps, covering the licensing (step 0), designing (step 1–5), implementation (steps 6–9), and presentation (steps 10–11) of the data in the semantic web. In the work presented here we follow these steps:

2.3.1 Licensing

WikiPathways content is covered by the Creative Commons Attribution 3.0 Unported license (https://creativecommons.org/licenses/by/3.0/). This is stated in the VoID headers of the RDF made. These headers are automatically generated by the same script generating the WikiPathways RDF. Open PHACTS provides a template for these header files.

2.3.2 Implementation

We used a Java RDF framework, Jena (http://jena.apache.org/), to generate the RDF for WikiPathways. The pathway diagrams were obtained through the web services of WikiPathways, after which they were converted into RDF with the Jena RDF framework. The code of the serializer is available on GitHub (https://github.com/wikipathways/wp2lod). The vocabularies were generated with a vocabulary framework called Deri Neologism (http://neologism.deri.ie/).

2.3.3 Presentation

The resulting RDF triples are available from (http://rdf.wikipathways.org) and loaded on a instance of the Virtuoso Open-Source Edition (http://virtuoso.openlinksw.com/) and available through its SPARQL endpoint at http://sparql.wikipathways.org. The triples are also loaded on the Open PHACTS discovery platform (https://dev.openphacts.org/docs/1.5) where they can be accessed through eleven API calls.

2.3.4 Identifier mapping

In the context of the semantic web, it is impractical to burden query writers with handling identifier mapping per resource and per query. Rather, the mapping results themselves need to become part of the semantic web. We applied two distinct approaches to addressing identifier mapping in our WikiPathways and Open PHACTS projects.

2.3.5 Query expansion

The Open PHACTS framework provides query expansion functionality through its Identifier Mappings Services. When an identifier is queried the SPARQL query is enriched with all possible identifiers to retrieve an expanded set of related entities. This approach is the most efficient in terms of the number of triples, since it requires only a single identifier per relationship, eliminating redundancy. However, it also requires a hosted identifier mapping service that it called along with every query.

2.3.6 Unified identifiers

In the case of WikiPathways, which does not host a mapping service, we chose a unified identifier approach, where all identifiers are mapped ahead of time to a set of common identifier systems. In this way, the database effectively contains the results of a limited number of identifier mappings in form of partially redundant triples. For example, in the WikiPathways RDF, all identifiers have been unified to Entrez Gene (wp:bdbEntrezGene), Ensembl (wp:bdbEnsembl), UniProt (wp:bdbUniprot) for gene products and HMDB (wp:bdbHmdb), and

ChemSpider (wp:bdbChemspider) for compounds like metabolites and drugs. The original identifier provided by the pathway curator is stored as a triple, with the predicate dc:identifier, and a URI from identifiers.org, which points to both the identifier and the resource.

2.4 Summary

We present a semantic web representation of WikiPathways together with vocabularies needed to cover the graphical pathway layout and the biological meaning and solutions to map between different identifier systems. The public availability allows rapid integration with other biological resources. The availability of two vocabularies allows to convert between different pathways resources. Different analytical tools now support the import of semantic web data, allowing integrated use of data from different resources with a single query. We demonstrate this with a federated query across multiple resources where the resulting differentially expressed genes for a disease where shown on a discovered pathway using PathVisio.

Availability

The following resources are publically available as beta releases just like WikiPathways. They are maintained as part of the open source WikiPathways project

Vocabularies

GPML: http://vocabularies.wikipathways.org/gpml
WP: http://vocabularies.wikipathways.org/wp

WikiPathways on the Semantic Web

SPARQL endpoint: http://sparql.wikipathways.org
Open PHACTS: https://dev.openphacts.org/docs/
RDF download: http://rdf.wikipathways.org

Source code

GitHub: https://github.com/wikipathways/wp2lod

References

1. A. Waagmeester *et al.* Using the Semantic Web for Rapid Integration of WikiPathways with Other Biological Online Data Resources. *PLOS Computational Biology*. 2016. **12** (6): 1–11.

2. D. G. Jennen *et al.* Biotransformation pathway maps in WikiPathways enable direct visualization of drug metabolism related expression changes. *Drug Discovery Today*. 2010. **15** (19-20): 851–858.

3. P. Khatri, M. Sirota, A. J. Butte. Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. *PLoS Computational Biology*. 2012. **8** (2): e1002375.

4. M. P. van Iersel *et al.* Presenting and exploring biological pathways with PathVisio. *BMC Bioinformatics*. 2008. **9** (1): 399.

5. T. Kelder, B. R. Conklin, C. T. Evelo, A. R. Pico. Finding the Right Questions: Exploratory Pathway Analysis to Enhance Biological Discovery in Large Datasets. *PLoS Biology*. 2010. **8** (8): e1000472.

6. T. Kelder *et al.* WikiPathways: building research communities on biological pathways. *Nucleic Acids Research*. 2011. **40** (D1): D1301–D1307.

7. M. Kutmon *et al.* WikiPathways: capturing the full diversity of pathway knowledge. *Nucleic Acids Research*. 2016. **44** (D1): D488–D494.

8. D. Maglott, J. Ostell, K. D. Pruitt, T. Tatusova. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research*. 2010. **39** (Database): D52–D57.

9. A. Yates *et al.* Ensembl 2016. *Nucleic Acids Research*. 2015. **44** (D1): D710–D716.

10. The UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Research*. 2014. **43** (D1): D204–D212.

11. D. S. Wishart *et al.* HMDB 3.0—The Human Metabolome Database in 2013. *Nucleic Acids Research*. 2012. **41** (D1): D801–D807.

12. H. E. Pence, A. Williams. ChemSpider: An Online Chemical Information Resource. *Journal of Chemical Education*. 2010. **87** (11): 1123–1124.

13. S. Kim *et al.* PubChem Substance and Compound databases. *Nucleic Acids Research*. 2015. **44** (D1): D1202–D1213.

14. A. P. Bento *et al.* The ChEMBL bioactivity database: an update. *Nucleic Acids Research*. 2013. **42** (D1): D1083–D1090.

15. M. P. van Iersel *et al.* The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. *BMC Bioinformatics*. 2010. **11** (1): 5.

16. *Semantic Web*, (http://www.w3.org/standards/semanticweb/).

17. T. Berners-Lee, R. Fielding, U. Irvine, L. Irvine, *RFC 2396 - Uniform Resource Identifiers (URI): Generic Syntax*, (http://www.faqs.org/rfcs/rfc2396.html).

18. N. Juty, N. L. Novere, C. Laibe. Identifiers.org and MIRIAM Registry: community resources to provide persistent identification. *Nucleic Acids Research*. 2011. **40** (D1): D580–D586.

19. A. Miles, S. Bechhofer, *SKOS Simple Knowledge Organization System Reference*, (http://www.w3.org/TR/skos-reference/).

20. J. Pinero *et al.* DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database*. 2015. **2015**: bav028–bav028.

21. J. S. Luciano. PAX of mind for pathway researchers. *Drug Discovery Today*. 2005. **10** (13): 937–942.

22. N. L. Novère *et al.* The Systems Biology Graphical Notation. *Nature Biotechnology*. 2009. **27** (8): 735–741.

23. K. W. Kohn, M. I. Aladjem, J. N. Weinstein, Y. Pommier. Molecular Interaction Maps of Bioregulatory Networks: A General Rubric for Systems Biology. *Molecular Biology of the Cell*. 2006. **17** (1): 1–13.

24. A. Finney, M. Hucka. Systems biology markup language: Level 2 and beyond. *Biochemical Society Transactions*. 2003. **31** (6): 1472–1473.

25. N. Juty *et al.* BioModels: Content, Features, Functionality, and Use. *CPT: Pharmacometrics & Systems Pharmacology*. 2015. **4** (2): 55–68.

26. *OWL 2 Web Ontology Language Document Overview (Second Edition)*, (http://www.w3.org/TR/owl2-overview/).

27. A. Luna, M. L. Sunshine, M. P. van Iersel, M. I. Aladjem, K. W. Kohn. PathVisio-MIM: PathVisio plugin for creating and editing Molecular Interaction Maps (MIMs). *Bioinformatics*. 2011. **27** (15): 2165–2166.

28. J. Ratnam *et al.* The Application of the Open Pharmacological Concepts Triple Store (Open PHACTS) to Support Drug Discovery Research. *PLoS ONE*. 2014. **9** (12): e115460.

29. H. Halpin, P. J. Hayes, J. P. McCusker, D. L. McGuinness, H. S. Thompson, *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, 2010, pp. 305–320, (https://doi.org/10.1007/978-3-642-17746-0_20).

30. E. Prud'hommeaux, A. Seaborne, *SPARQL Query Language for RDF*, 2008, (https://www.w3.org/TR/rdf-sparql-query/).

31. *Help:WikiPathways SPARQL queries*, 2015, (http://www.wikipathways.org/index.php/Help:WikiPathways_Sparql_queries).

32. van Hage, *SPARQL for R Tutorial - Linked Open Piracy*, 2015, (https://semanticweb.cs.vu.nl/R/sparql_lop/sparql_lop.html).

33. A. Bohler *et al.* Automatically visualise and analyse data on pathways using PathVisioRPC from any programming environment. *BMC Bioinformatics*. 2015. **16** (1): 267.

34. A. Waagmeester, *Differentially expressed Genes in a pathway on Insulin Signaling in case of Diabetes Mellitus*, 2015, (https://gist.github.com/andrawaag/6989c8c218862a912ef6).

35. E. G. Cerami *et al.* Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Research*. 2010. **39** (Database): D685–D690.

36. A. Waagmeester, *andrawaag/BioSystems2RDF*, 2015, (https://github.com/andrawaag/BioSystems2RDF).

37. A. J. Williams *et al.* Open PHACTS: semantic interoperability for drug discovery. *Drug Discovery Today*. 2012. **17** (21-22): 1188–1198.

38. C. Haupt, A. Waagmeester, M. Zimmermann, E. Willighagen, (http://www.openphacts.org/specs/2013/WD-rdfguide-20131007/).

39. B. McBride. Jena: a semantic Web toolkit. *IEEE Internet Computing*. 2002. **6** (6): 55–59.

3

Understanding signaling and metabolic paths using semantified and harmonized information about biological interactions

Adapted from: R. A. Miller *et al*. Understanding signaling and metabolic paths using semantified and harmonized information about biological interactions. *PLOS ONE*. 2022. **17** (4): e0263057.

Abstract

To grasp the complexity of biological processes, the biological knowledge is often translated into schematic diagrams of, for example, signalling and metabolic pathways. These pathway diagrams describe relevant connections between biological entities and incorporate domain knowledge in a visual format making it easier for humans to interpret. Still, these diagrams can be represented in machine readable formats, as done in the KEGG, Reactome, and WikiPathways databases. However, while humans are good at interpreting the message of the creators of diagrams, algorithms struggle when the diversity in drawing approaches increases. WikiPathways supports multiple drawing styles which need harmonizing to offer semantically enriched access. Particularly challenging, here, are the interactions between the biological entities that underlie the biological causality. These interactions provide information about the biological process (metabolic conversion, inhibition, etc.), the direction, and the participating entities. Availability of the interactions in a semantic and harmonized format is essential for searching the full network of biological interactions. We here study how the graphically-modelled biological knowledge in diagrams can be semantified and harmonized, and exemplify how the resulting data is used to programmatically answer biological questions. We find that we can translate graphically modelled knowledge to a sufficient degree into a semantic model and discuss some of the current limitations. We then use this to show that reproducible notebooks can be used to explore up- and downstream targets of MECP2 and to analyse the sphingolipid metabolism. Our results demonstrate that most of the graphical biological knowledge from WikiPathways is modelled into the semantic layer with the semantic information intact and connectivity information preserved. Being able to evaluate how biological elements affect each other is useful and allows, for example, the identification of up or downstream targets that will have a similar effect when modified.

Author summary

Resources like WikiPathways contain many biological pathway diagrams and within these diagrams are even more pathway elements, representing genes, proteins, and metabolites. In the case of WikiPathways, the basic elements of the diagrams are nodes with biological information about gene products, metabolites, as well as other pathways, and edges that represent the biological interactions, complemented with graphical elements meant to make diagrams easier to read. While these elements can generally be understood by a biologist upon visual inspection, it takes implementation of technologies like RDF and shape expressions (ShEx) to make the pathway diagrams able to be batch queried by a computer. This allows researchers to query the entire resource at once to observe systemic patterns. The work presented here is intended to inform how biological elements are interacting with one another and how to leverage this information to answer biological questions.

3.1 Introduction

Human cells contain around 20,000 protein-coding genes and numerous non-coding genes (*2*) and each coding gene can encode many proteins. Furthermore, the Human Metabolome Database (HMDB) describes over 100,000 metabolites (*3*). The number of interactions between biological entities is even higher. For example, cells also contain many membrane and soluble protein complexes (*4*), the latter estimated as at least 600 (*5*), while many more are predicted (*6*). The size and complexity of the system gives a system-wide overview, but sometimes breaking the system into smaller pieces that can be used for analysis and experimentation is wanted (*7*, *8*).

WikiPathways is an open source pathway repository that is open to the community to create and modify pathway diagrams so that they can be shared with everyone in the community (*9*). The WikiPathways database depicts biological processes and their connections to each other. The connections of elements within a pathway are shown as edges from one node to the next. These edges themselves have bi-

Table 3.1: **Abbreviations for semantic web technologies used to harmonize the biological interaction information from WikiPathways.**

Abbreviation	Full Name/Meaning
GPML	Graphical Pathway Markup Language
GPMLRDF	RDF for Graphical Pathway Markup Language
MIM	Molecular Interaction Map
RDF	Resource Description Framework
SBGN	Systems Biology Graphical Notation
ShEx	Shape Expressions
SPARQL	SPARQL Protocol and RDF Query Language
WikiPathways RDF	The combination of GPMLRDF and WPRDF
WPRDF	RDF for WikiPathways

ological meaning that can be modelled and represented in WikiPathways (*10*).

For interoperability, WikiPathways also has a Resource Description Framework (RDF) set associated with it (*11*). The RDF is the semantic representation of pathway diagram elements that are displayed and generated from the original Graphical Pathway Markup Language (GPML) in which WikiPathways stores the pathways (see Table 3.1 for terminology used in this article). The WikiPathways RDF then includes both the graphical RDF (GPMLRDF) and the semantic elements of the RDF (WPRDF). The RDF allows users to go from creating an image of a biological pathway to trapping the elements and keeping them in a machine readable way and made available to be queried. One of the advantages of this is that it is also a linked data resource that can be queried by users at the WikiPathways SPARQL Protocol and RDF Query Language (SPARQL) endpoint, to query RDF databases (http://sparql.wikipathways.org/). This store of the WikiPathways RDF can be accessed both directly from the WikiPathways SPARQL endpoint, but also by remote requests via federated queries.

In order to represent connectivity between nodes in a pathway diagram, the meaning of a drawn line connecting nodes needs to be understood. WikiPathways RDF has connectivity information stored as point A is connected to point B. To a human looking at a pathway, it is more obvious what an arrow connecting two points means or what is implied by the arrow, but the RDF needs this stated explicitly if any inferences about how elements are connected is to be gleaned. In fact that is even true when standardised graphical representations for interactions like Molecular Interaction Maps (MIM) (*12*) and Systems Biology Graphical Notations (SBGN) (*13*) are used.

Furthermore, to ensure the biological causality is reflected in the graph representation in the RDF, we need to make sure the latter reflects that interactions can be directed and undirected. Information about the direction and connectivity in a pathway diagram helps to explain the biological processes and therefore helps understand cause-effect relationships represented in the pathway. However, not all interactions have a clear direction: while the direction of a metabolic conversion follows chemical thermodynamics, interactions like the associations that exist in a complex are symmetrical and do not have a direction. Even more complex is a ligand binding, where the physical interaction is not only directed, but the interaction arrow also reflects the movement of the ligand. Therefore, it is important to know if an interaction has a directed route as part of a path and the RDF needs to preserve this information.

To ensure that pathway interaction drawings and notations can be biologically interpreted, the RDF needs to have standardized types for the interaction. That will allow users to query for all reactions of a similar (biological) type rather than worry about which notation was used in the drawing. WikiPathways supports several drawing notations, which can be general WikiPathways notations, MIM notations, and SBGN notations. Based upon WikiPathways GPML data model and the underlying ontology, these three can all be used and shown on WikiPathways. The available interactions themselves can be classified into nine different types: conversions, bindings, interactions, directed interactions, catalysis, transcription translation, complex bindings, inhibitions, and stimulations.

When interactions in various notations are normalized, more biological knowledge can be explored, and new questions answered. This interoperability effort makes it possible to gain implied knowledge from how a pathway diagram is drawn. For example, if two enzymes are catalyzing some chemical substrates in succession then there would typically not be a direct link or arrow drawn from one enzyme to the other, but in order for the second enzyme to work the product from the first reaction must be present. This has the implication that the second enzyme is biologically downstream of the first enzyme, even though this interaction is not explicitly drawn. Having semantically clear directions and interaction types is essential to reach this conclusion from the RDF. Drawing of interactions with the WikiPath-

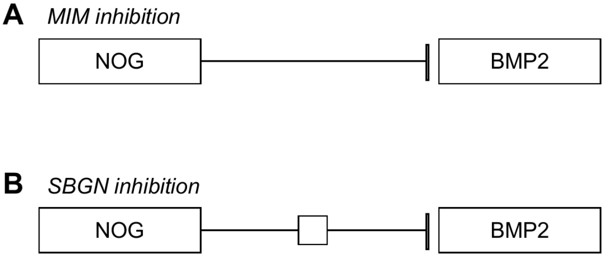


Figure 3.1: Differences in drawing of MIM vs SBGN inhibition interaction. A shows a MIM - inhibition interaction. B shows a SBGN - inhibition interaction.

ways and MIM notations can be done with the default installation of the PathVisio core (*10*), while SBGN needs a PathVisio plugin https://github.com/PathVisio/pathvisio.github.io/blob/master/plugins/sbgn.md.

The PathVisio pathway editor thus makes it possible to annotate an interaction as a simple line with an arrowhead, as a MIM interaction, by default, or to create a SBGN drawing using plugins. It then becomes necessary to unify common types from the different graphical standards so that a MIM-Inhibition and a SBGN-Inhibition are understood as the same thing. Figure 3.1 shows the differences in drawing of an inhibition between SBGN and MIM notation. After all, in both cases, the interaction is indicating an inhibitory effect of one entity upon another. Knowing the interaction types gives important context of the connection and the entities involved. A small note about how complexes are represented is also essential. In the RDF all the entities are connected to each other with an undirected interaction. This keeps them all connected to each other as well as with any interaction that they are associated with as a complex.

The general interaction type is used to denote an interaction between data nodes and thus all interactions are of this type. A directed interaction, on the other hand, means there is a direction that says one data node is influencing another but the exact mechanism is not known, or at least not described by the pathway creator (author). Directed interaction is also the general data type for all interactions that have some directional information included. Therefore, all interactions have the type directed interaction except binding and complex binding, with the directed interaction itself being a child of the general interaction type. We therefore wanted to study to what extent we can derive knowledge from biological interactions, by semantically capturing biological meaning of interactions and harmonizing the notation in pathway drawings. We tested our hypothesis that this can be done by answering the following questions. First, can we translate graphically modelled biological knowledge to a semantic model of biological knowledge that harmonizes interaction types and captures implied directional information And second, can we then take advantage of the semantic translation of the graphical biological knowledge to programmatically answer biological questions. For this latter question, we studied two specific biological questions as examples: in one example we look at MECP2 and explore alternative targets for this protein by looking for targets either upstream or downstream as they both have an effect on MECP2's role. For the other example we studied how lipid metabolism is captured in the *Ganglio Sphingolipid Metabolism* pathway (wikipathways:WP1423, WikiPathways Project *et al.*, 2019).

The description of interaction information allows for the advancement of curation efforts by the WikiPathways team. This curation in turn allows the team to improve the quality of pathways and a more complete overview of which elements are in the pathways and how they are connected to one another. Using SPARQL queries for curation the curators can identify why the interactions are not converted from the graphical description of WikiPathways to the semantic description of the WikiPathways RDF and can explore how to improve this.

When we understand how interactions work we can also pre-define the form or shape that such a specific interaction type takes. For this the Shape Expressions (ShEx) standard can be used (*14–16*). A ShEx determines what information is expected for, in this case, a specific interaction type. ShEx will be created for all interaction types in WikiPathways. The shape expression can then be used to monitor translations of knowledge of one format or notation to another, for example, when adding data from one database to another (*17*). This allows us to focus more on the biology and less on the bioinformatics, as we get alerted about unexpected shapes.

To explore these approaches, we look at two biological research topics studied in our group: a rare disease and human lipid metabolism. MECP2 is a protein involved in a rare disease and important in the methylation of DNA (*18*). Mutations in the MECP2 gene have been linked to the development of Rett Syndrome (*19*). This disease is responsible for a host of neurological developmental issues that affects infant development. The MECP2 gene lays on the X-chromosome and Rett Syndrome is found in females (*20*) because the severity in males is too high for patients to be viable. The severity of the disorder is related to the specific mutation found in the individual patient (*21*). Ehrhart *et al.* have already demonstrated the power of integrating different databases to retrieve links between genetic variants and phenotypes (*22*). Being able to look at alternative targets that are a part of the sequence of developments that lead to disorders such as Rett may end up helping us to expand the knowledge about alternative causes and treatment opportunities. The types of interactions described for MECP2 are a simple case of connectivity and directional information captured in WikiPathways and make a good example to demonstrate how this can be used to allow observation of upstream and downstream interactions.

The second example describes the metabolic regulation and modifications of sphingolipids which are known to regulate several cell functions (*23*). Sphingolipids are produced in the endoplasmic reticulum and the modifications of this lipid class alters the effect of the specific sphingolipid's function (*24*). The conversion of these metabolites from one form to another is regulated by enzymes that act as a catalyst for the reaction to take place. Sphingolipids also play a role in signal transduction (*25*). The sphingolipids play an important role in the membrane of eukaryotic cells and are often associated with disorders in the degradation of lipids (*26*). This shows the importance of proper metabolite regulation and metabolism as disruptions can lead to serious diseases with high mortality rates. Understanding how these elements of the pathway are connected to one another and how they are directed helps to understand when the elements are not working correctly. There are also a large number of proteins that are known to interact directly with sphingolipids and are necessary for cell function (*27*). In WikiPathways, these types of interactions are most often drawn with an arrow that shows the conversion of the metabolites from one form to another along with an associated catalysis reaction that is facilitated by an enzyme. Looking at how metabolism is modelled in wikipathways:WP1423 helps illustrate how these conversion and catalysis reactions are stored. Metabolism interactions are a more complicated set of interactions as an enzyme is typically seen acting on another interaction. The sphingolipid metabolism pathway displays this more complex observation and allows the identification of the order of the enzymes found for potential upstream/downstream analysis.

3.2 Materials and methods

WikiPathways Data

Interaction modeling

The interactions in WikiPathways are modeled by taking the graphical semantic information from the pathway diagram's GPML representation. The harmonization of interactions is part of the WPRDF generation. This is done by analysis of the lines that represent interactions in the graphical representation, and using these to decide how the participants in the interactions are connected. All harmonized interactions have a unique ID, are linked to the participants, and have an interaction type as outlined in the introduction. If it is a directed interaction, it will also have a source and target node for the interaction. JUnit (https://junit.org/) was used to test the harmonization with several tests to verify that these connections in the GPML are being converted to RDF as expected. These tests include the original GPML and the expected outcomes as described in the code repository at https://github.com/BiGCAT-UM/WikiPathwaysInteractions/tree/master/FilesGPML.

Benchmark data

We used the RDF from the WikiPathways June 2019 release (https://zenodo.org/record/3369380). Both the WPRDF and the GPMLRDF components of the WikiPathways RDF were used in this study. To examine how pathways are drawn and used in WikiPathways, the analysis used only pathways from the Curated collection and only for *Homo sapiens*, and therefore excludes the Reactome collection (*28*, *29*).

Data Analysis

To aggregate and analyze the date, Jupyter Notebooks running Python were used to collect all SPARQL queries that were used to query the WikiPathways SPARQL endpoint (*30*). The notebooks are available from (https://github.com/BiGCAT-UM/WikiPathwaysInteractions/): *DataNodeStats.ipynb*, and *InteractionStats.ipynb*, and two for the two biological examples. The first two represent two different categories of queries. *DataNodeStats* retrieves information about data nodes in both parts of the WPRDF while the *InteractionStats.ipynb* file is used to return data about connectivity between the nodes in the WikiPathways RDF, representing both the semantic and the graphical RDF elements. *ExampleMECP2.ipynb* is the file for the query related specifically to the *MECP2 up and down stream targets* example. Finally, *ExampleLipidMetabolism.ipynb* is the notebook for the case of *sphingolipid metabolism*. These notebooks and their use are further described below.

Datanode Harmonization

Data nodes needed to be harmonized first in order to be able to examine the connections between the nodes. There are two conditions that determine the conversion of the interactions: the participating datanodes are converted, and second, the interaction is converted. That allowed us to better estimate how well the interaction harmonization itself went. Therefore, we first looked at the data nodes. The *DataNodeStats.ipynb* notebook contains Python code to calculate a series of counts of data nodes, to estimate the amount of data and to get a baseline number of what we can expect for the success of conversion and harmonization of interactions. It is important to realize that for interactions where one of the participating data nodes is not in the WPRDF, the conversion script will not to be able to create the interaction due to the absence of participants. Therefore this interaction will not be found in the WPRDF and will affect our interaction counting. The notebook calculates the total number of data nodes of a certain type, in the Jupyter Notebook section *Datanode Type Counts*, and the corresponding numbers of GPMLRDF data nodes without a WPRDF data node equivalent. Furthermore, it determines the number of GPML-RDF data nodes of type complex without WPRDF equivalents. This is used to specifically track which data nodes that are part of the complexes that can be found in the graphical elements part of the RDF but not found in the WPRDF, the biological component of the WikiPathways RDF. These complexes are not annotated as biologically known complexes. Those exist because the biological meaning of complexes is currently not always well-defined in pathway drawings in WikiPathways.

Interaction Harmonization

The *InteractionStats.ipynb* notebook contains code to calculate numbers that reflect the harmonization of interactions in the biological WPRDF, by taking into account the different drawing notations as a unified interaction type. The first few sections calculate overall statistics, the *Number of Non-Directed Interactions* (for example, bi-directional binding), *Count of Interaction Types* (reflecting the biological nature of the interaction), *Interaction Count with Unspecified Type*, and the percentage of non-directed interactions. The second set of sections characterize the nature of the interactions, e.g. *Interaction counts by participants*, *Participants for Interactions* (which reflects what datanode types are involved in an interaction), and *Identifier IDs by data source*.

In order to evaluate the conversion success, it calculates the complementary *GPMLRDF Interactions without a WPRDF equivalent* and *GPMLRDF Interactions with a WPRDF equivalent*, and the resulting percentages of success (see *GPMLRDF Interaction with Equivalent WPRDF out of Total GPMLRDF Interactions*). The GPMLRDF Interactions without a WPRDF equivalent was used to check to see how many interactions that are present in the graphical version of the RDF but not present in the biological WPRDF. The query for the percentage of WPRDF Interactions that are of unspecified type was used to see how accurately detailed the biological pathways are annotated. Finally, the percentage of non-directed interactions in the notebook calculated how many of the WikiPathways interactions are of non-directed type. When these are between metabolites and they may reflect missing biological annotation of directions.

Usability

To test our hypothesis that we can harmonize the interaction information, we developed the Jupyter Notebooks to first collect and query the data from the WikiPathways RDF. We then created several unit tests to validate how the modelled interactions behaved and to verify that they are created correctly. This ensures that when an interaction is drawn, we can keep track of the relevant semantic data represented, such as what nodes are connected to each other, what type of interaction is drawn between them, and how many nodes are expected to be part of the interaction. We can then test assumptions like: "interactions between metabolites should be directed conversions" and "interactions between different proteins should not be conversions" and add other aberrant results as curation tasks. We further tested with two biological examples if the harmonized semantified interactions give interpretable answers.

Curation

The Jupyter Notebook created for interaction curation uses the query for GPML RDF interactions without a WP RDF equivalent to generate a list of interactions that are not found in the semantic portion of the RDF. The next query in the notebook finds the specific elements for the interactions in this list that will help the curator identify which elements are missing. The query includes the interaction ID for the GPML RDF, the pathway in which it can be found, and the connecting elements found on either end of the interacting line.

ShEx

Shape expressions were created manually for the modelled WikiPathways interactions. ShEx for WikiPathways interactions were formed following the standards laid out by the ShEx project (https://shex.io/). These shape expressions can be found in the shape expressions subdirectory on the GitHub repository (https://github.com/BiGCAT-UM/WikiPathwaysInteractions/tree/master/ShExInteractions). The harmonized interaction types were expressed as ShEx. ShEx can be used for curation events to verify that the interaction fits the shape that is expected by the WikiPathways model, and in this way help detect data issues. The npm module shex (https://www.npmjs.com/package/shex) was used to run the shape expression on the harmonized model. A GNU/Linux Makefile on GitHub demonstrates the combination of SPARQL to list all resource IRIs of a certain interaction type and the JSON query tool jq (https://stedolan.github.io/jq/) to process the ShEx module output to count the number of errors for each interaction. This allowed running the shape expression on all directed interaction in the WPRDF.

MECP2 up- and downstream targets

For the specific example used for MECP2 metabolism, the Jupyter Notebook used a SPARQL query to the WPRDF. This query works by first searching for targets that are upstream or downstream of MECP2. The query then identifies data nodes that are associated with the HGNC symbol MECP2. The query in the Jupyter Notebook finally finds associated pathways that have this HGNC symbol present and matches interactions that have MECP2 as a target in the interaction.

Sphingolipid metabolism

In the case of the specific example used for sphingolipid metabolism, the Jupyter Notebook used a SPARQL query to the WPRDF. The query retrieves the source portion of an interaction and displays its label. In the case of sphingolipid metabolism, the queries identified enzymes that are associated with conversions in the pathway and returned results with the enzyme, interaction, the source metabolite and the target metabolite product.

Table 3.2: **Datanode Type Counts, as defined by the WikiPathways ontology. The Datanode counts for each type of node.**

Datanode Type	**Count (WPRDF)**	**Count (GPMLRDF but not WPRDF)**
Datanode	28402	—
GeneProduct	21270	1084
Protein	8255	141
Metabolite	4038	219
RNA	1204	66
Complex	980	16
Unknown	—	218
Pathway	—	250

3.3 Results

To understand the amount of data that can be accessed via the RDF, we looked at the available RDF data for WikiPathways as GPMLRDF and WPRDF, the first being a direct translation of the original graphical depiction of the GPML files and the second covering the biological content. A quick count of the June 2019 release shows that the WPRDF used in this paper had 24,220 data nodes, and 13,928 interactions and is available at http://data.wikipathways.org. The subject of the paper is the interactions between data nodes, but we first need to understand that edges of a network connect datanodes to one another and so understanding the fundamentals of the biomolecular data nodes is necessary. This defines some context for the following results.

Datanode Results

With regards to the data nodes, because of the hierarchical annotation the most prevalent node type is the general datanode type. It is the base type for any datanode, as described by the WikiPathways Vocabularies (https://vocabularies.wikipathways.org) and thus is used for every data node, it may include any of the descriptive data types. More specific but still generic, the GeneProduct type is the next most prevalent node type. These include explicitly typed proteins and RNAs and while the remaining GeneProduct typed nodes are not specified further. Table 3.2 illustrates the size of the WikiPathways semantic RDF part and the types of nodes present in WikiPathways. There are a total of 28,402 data nodes, the majority of which are gene products. Proteins are the next common type followed by metabolites and RNA. There are also Complex nodes to represent clustered groups of other node types, specifically proteins, gene products, and RNA. Pathways are not typed as Datanode in the WPRDF, which is why the value is blank in the table. Overall, 7.0% of GPMLRDF data nodes do not have a WPRDF data node equivalent and thus 93.0% of the GPML data nodes are found in both parts of the RDF.

Also seen in Table 3.2 are the data nodes that are found in the GPMLRDF but not found in the WPRDF. The reason typically is that the node exists but is not linked to a clear biomolecular database identifier, in other words we do not know exactly what it is. Datanodes are any node type in the pathway diagram and the count of gene products also includes proteins and RNAs as these are specifications of the products produced. Complexes are a combination of several other node types that form a unit with one another. We can also see how many data nodes are found in both parts of the RDF.

If we specifically look at some examples of data nodes that are present in the GPMLRDF but not carried over to WPRDF, we can see a list of sixteen complex data nodes, and the details of these are given in the S5 File. This second table also includes the labels for the complexes, shedding some light on which complexes were not transferred over to the semantic portion (WPRDF) of the RDF from the graphical portion (GPMLRDF). For all these nodes, they lacked database identifiers.

When we do this evaluation for the pathways of the two use cases, we find that for wikipathways:WP4312, which pertains to MECP2, there is 1 gene product type data node that is found in the GPMLRDF but not found in the WPRDF. This represents 1 gene product out of 148 other gene products that were found in the WPRDF and out of 152 total data nodes found in the WPRDF. In the instance for wikipathways:WP1423, which is related to sphingolipid metabolism, there is 1 metabolite that is found in the GPMLRDF but is not found in the WPRDF. This is 1 metabolite from 38 total metabolites found in the sphingolipid metabolism pathway and out of 62 data nodes found in the WPRDF for wikipathways:WP1423. The last metabolite (Gal-GlcNAc-GM1b) is modified with two sugars, and not found in the reference databases. Future WikiPathways releases can annotate such nodes with the InChIKey, for which no database record is required.

In the S1 File there are tables with examples of data node types that are found in the GPMLRDF but not in the WPRDF for various pathways (as counted in Table 3.2). In this file, the query results are retrieved along with the table to give some idea why they may not be translated. In the S2 and S3 Files there are tables for the data node counts for the specific WikiPathways example pathways of MECP2 and sphingolipid metabolism.

Interaction Results

Similar to what we did for the data nodes, we calculated non-directed interactions and non-specific interactions along with the specific interaction types and counts. Non-directed interactions being all interactions that do not have any directional information, such as in the case of a binding event. Non-specific, on the other hand, means that an interaction does not even have a specified non-directed interaction like a binding.

First, we identified nine interaction types. The overview of mappings to WPRDF of the GPML interaction types that can be found in WikiPathways, is available from https://github.com/BiGCaT-UM/WikiPathwaysInteractions/tree/master/FilesGPML. The nine types of interactions found in the GitHub page are catalysis, complex binding, conversion, general undirected interaction, inhibition, stimulation, transcription/translation, an

Table 3.3: **Interaction Type Counts, as defined in the WikiPathways ontology.** The sum of DirectedInteration and NonDirected equals the Interaction Total. Of the directed interactions, subsets are typed as Conversion, Inhibition, etc. The NonSpecified interactions is a subset of NonDirected interactions. More than 12 thousand interactions are only found in the GPMLRDF.

Interaction Type	**Count (WPRDF)**	**Count (GPMLRDF but not WPRDF)**
Interaction	15525	—
DirectedInteraction	11819	—
Conversion	1447	—
Inhibition	1091	—
Catalysis	1231	—
ComplexBinding	940	—
Binding	1513	—
Stimulation	842	—
TranscriptionTranslation	256	—
NonDirected	3706	—
NonSpecified	2766	—
Unknown	—	12287

unspecified directed interaction, and a directed interaction with multiple inputs and multiple outputs. This GitHub repository contains example GPML files for each interaction type that can be found at https://vocabularies.wikipathways.org/, along with an example of what the interactions look like in GPML, as well as files with statistics about the interaction as it appears in the WPRDF. These numbers are used in the JUnit tests to verify that the different models are harmonized into the single interaction model in WPRDF. These tests are now available as part of the regular testing of RDF generation (see https://github.com/wikipathways/GPML2RDF, *src/test/java/org/wikipathways/wp2rdf/interactionTests* folder). When we look at the full WPRDF, the types of generic non-directed and nonspecific interactions can be seen. Out of a total of 15,525 interactions, 3,706 (23.9%) were non-directed of which 2,766 (17.8%) were non-specific (see Table 3.3). Thus 11,819 (58.3%) of the interactions have some sort of direction information. The number of non-specific interactions can be either an indication that there is just not sufficient evidence to explain what the interactions are or that better curation is necessary. Examples of how interactions are drawn in WikiPathways can be seen in Figure 3.2.

Only a small percentage of the interactions have associated identifiers. Having such identifiers can make it easier to find information about the provenance of that interaction occurring in a pathway and it is useful for linking experimental data or modelling results to the pathway or to find descriptions of the interactions in external resources. Table 3.4 contains provenance information about the databases to which identifiers for interactions refer. UniProt-TrEMBL has the most interactions represented in WikiPathways. There were some unexpected database links. Sources like *kegg.compound* and *ChEBI* are not expected to have interaction data information but are included because the user identified them as the database resource for the interaction. These unexpected sources come from two pathways, wikipathways:WP3634, and wikipathways:WP3635. These two pathways use very specific notation and while unexpected, have been intentionally annotated like this. These pathways use the SBML notation and represent the normal versus disease state of insulin signaling (*31*). Generally, the main reason that currently most interactions do not have any database identifier associated with them is that the mechanism to add these is relatively new.

Finally, to further characterize the interactions present, Tables 3.5 and 3.6 provide examples of the makeup of the interactions seen in WikiPathways. Table 3.5 shows example Interaction IDs, along with their interaction types, and what type of datanode type is participating in the interaction. And Table 3.6 shows the profile with the interaction participants and a count of how many times this interaction profile was counted in WikiPathways and the type of these interactions.

When the *PathwayStatsMECP2.ipynb* and *PathwayStatsSphingolipid.ipynb* notebooks were applied to the pathways of the two use cases, we found

Table 3.4: **Interaction Identifier ID counts by data source.**

Database Source	Interactions
Rhea	313
Uniprot-TrEMBL	213
KEGG Pathway	28
pato	8
kegg.compound	8
ChEBI	6
KEGG Reaction	3
Reactome	3
WikiPathways	2
XMetDB	2
SPIKE	2
BIND	1

that for wikipathways:WP4312, which pertains to MECP2, there are 5 interactions that are found in the graphical GPMLRDF but not found in the semantic WPRDF. This represents 5 interactions out of 45 non-specified interactions that were found in the WPRDF and out of 37 directed interactions found in the WPRDF. In the instance for wikipathways:WP1423, which is related to sphingolipid metabolism, there are 24 interactions that are found in the GPMLRDF but not found in the WPRDF. Still, we find 49 directed interactions in the WPRDF for the sphingolipid metabolism pathway, of which 13 are typed as catalytic reactions.

In the S2 and S3 File tables can be found for the interaction counts of the two specific pathways for MECP2 and sphingolipid metabolism. These contain the types of interactions found in these pathways as well as how many interactions were found in the GPMLRDF but not in the WPRDF resources for WikiPathways as described above.

Table 3.5: **Participants for Interactions. Twenty example interaction syntaxes shown in table below.** First twenty interactions from the WikiPathways RDF along with their interaction type and the participants for each interaction

Interaction	Interaction Type	Interaction Participants
WP3668_r97639/ComplexBinding/b916e	Binding	Complex, GeneProduct
WP2879_r94789/ComplexBinding/c939e	Binding	Complex, GeneProduct, Metabolite
WP4262_r97132/ComplexBinding/dae4b	Binding	Complex, GeneProduct, Metabolite
WP585_r94686/WP/Interaction/ida141949	Catalysis	GeneProduct, Protein
WP2533_r95594/WP/Interaction/adbe3	Catalysis	Conversion, DirectedInteraction, Interaction, Protein
WP1601_r95004/WP/Interaction/ida833b0dc	Catalysis	Conversion, DirectedInteraction, GeneProduct, Interaction
WP1423_r94289/WP/Interaction/idde73da53	Catalysis	DirectedInteraction, GeneProduct, Interaction
WP3865_r88186/ComplexBinding/d5e4f	ComplexBinding	Complex, GeneProduct
WP2446_r87639/ComplexBinding/e75ff	ComplexBinding	Complex, GeneProduct, Protein, Rna
WP2795_r97631/ComplexBinding/b5fa4	ComplexBinding	Complex, GeneProduct, Protein
WP3580_r96434/WP/Interaction/id6d378f23	Conversion	Metabolite
WP134_r94935/WP/Interaction/a5dec	Conversion	Metabolite
WP3627_r90137/WP/Interaction/id14d637fe	Conversion	Metabolite
WP2436_r97673/WP/Interaction/b1b2f	Conversion	Metabolite
WP4149_r94399/WP/Interaction/id30000f59	Inhibition	GeneProduct, Protein
WP2261_r89520/WP/Interaction/id65877034	Inhibition	GeneProduct, Protein
WP306_r97459/WP/Interaction/e8847	Inhibition	GeneProduct, Protein
WP2526_r96312/WP/Interaction/ddfe1	Stimulation	Protein
WP1984_r95143/WP/Interaction/id8ba5f251	Stimulation	GeneProduct, Metabolite
WP1984_r95143/WP/Interaction/iddde89331	Stimulation	GeneProduct, Protein

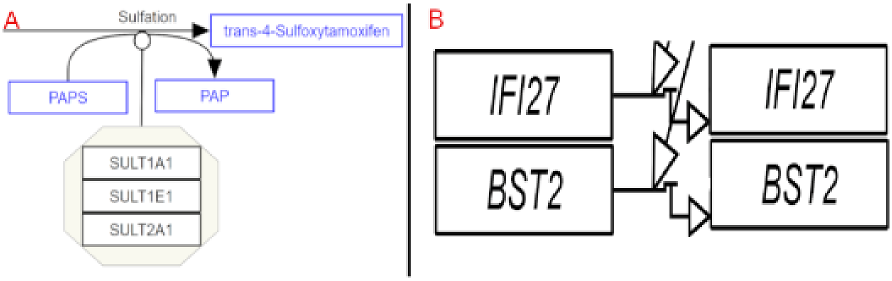


Figure 3.2: Interaction types that are not found in Table 6. A shows a complex binding of SULT1A1, SULT1E1 and SULT2A1 that catalyzes cis-4-hydroxytamoxafin to trans-4-sulfoxytamoxifen with PAPS to PAP formation found in Tamoxifen Metabolism (wikipathways:WP691). B shows transcription translation interaction for BST2 to BST2 in Host-pathogen interaction of human corona viruses - MAPK signaling pathway (wikipathways:WP4877).

Curation

As can be seen in the Jupyter Notebooks for curation, 11081 interactions are found in the GPMLRDF but not found in the WPRDF. The details for the first 20 results are found in Table 3.7. As can be seen in the table, the query identifies the interaction information from the GPMLRDF, the graph reference ID from the GPMLRDF, and the label for the participants. This can be used to help identify problematic interactions that are not being converted to the WPRDF.

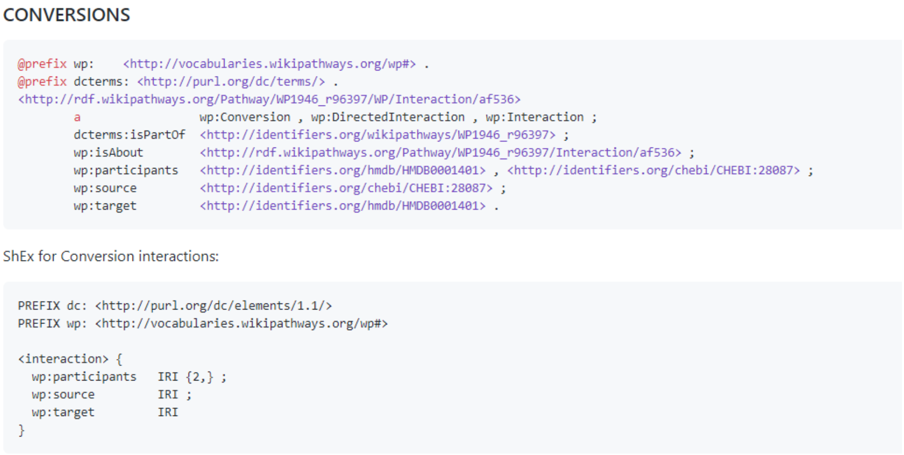
ShEx

All of the ShEx forms can be found on the GitHub repository https://github.com/BiGCAT-UM/WikiPathwaysInteractions/tree/master/ShExInteractions. The interaction types found at https://vocabularies.wikipathways.org/ are general WikiPathways interactions (wp:Interaction), the general WikiPathways directed interactions (wp:DirectedInteraction), the harmonized WikiPathways binding (wp:Binding), complex binding (wp:ComplexBinding), coversions (wp:Conversion), inhibitions (wp:Inhibition),

Table 3.6: **Top 20 most occurring directional interactions by participants combination. The most abundant interaction is a directed interaction between two metabolites.**

Interaction Participants	Count	Type
Metabolite, Metabolite	2675	DirectedInteraction
GeneProduct, GeneProduct	1423	DirectedInteraction
GeneProduct, Protein, GeneProduct, Protein	1334	DirectedInteraction
Metabolite, Metabolite	1125	Conversion
Metabolite	474	DirectedInteraction
GeneProduct, Protein, GeneProduct	445	DirectedInteraction
GeneProduct, GeneProduct, Protein	438	DirectedInteraction
GeneProduct, Protein	420	DirectedInteraction
GeneProduct	315	DirectedInteraction
DirectedInteraction, Interaction, GeneProduct	315	DirectedInteraction
GeneProduct, Protein, Protein	292	DirectedInteraction
Metabolite, GeneProduct	291	DirectedInteraction
DirectedInteraction, Interaction, GeneProduct	274	Catalysis
Protein, Protein	273	Stimulation
GeneProduct, GeneProduct	270	Inhibition
Protein, Protein	262	DirectedInteraction
DirectedInteraction, Interaction, Conversion, Protein	227	DirectedInteraction
DirectedInteraction, Interaction, Conversion, Protein	226	Catalysis
GeneProduct, Metabolite	180	DirectedInteraction
GeneProduct, DirectedInteraction, Interaction	151	DirectedInteraction

catalysis (wp:Catalysis), stimulations (wp:Stimulation), and transcription-translation (wp:TranscriptionTranslation) interactions. For example, the shape expression representation for a conversion interaction is seen in Figure 3.3. These represent the harmonized interaction types found in the WikiPathways RDF and their expression in ShEx.



CONVERSIONS

```
@prefix wp:    <http://vocabularies.wikipathways.org/wp#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
<http://rdf.wikipathways.org/Pathway/WP1946_r96397/WP/Interaction/af536>
        a                 wp:Conversion , wp:DirectedInteraction , wp:Interaction ;
        dcterms:isPartOf  <http://identifiers.org/wikipathways/WP1946_r96397> ;
        wp:isAbout        <http://rdf.wikipathways.org/Pathway/WP1946_r96397/Interaction/af536> ;
        wp:participants   <http://identifiers.org/hmdb/HMDB0001401> , <http://identifiers.org/chebi/CHEBI:28087> ;
        wp:source         <http://identifiers.org/chebi/CHEBI:28087> ;
        wp:target         <http://identifiers.org/hmdb/HMDB0001401> .
```

ShEx for Conversion interactions:

```
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX wp: <http://vocabularies.wikipathways.org/wp#>

<interaction> {
  wp:participants   IRI {2,} ;
  wp:source         IRI ;
  wp:target         IRI
}
```

Figure 3.3: Example ShEx shape for the WikiPathways harmonized Conversion interaction element (RDF shown in the top half), that requires two or more participant IRIs and exactly one source IRI and one target IRI.

MECP2 up and down stream targets

We created Jupyter Notebooks to evaluate the example pathways, as described in the Methods section. The SPARQL queries used in the Jupyter Notebooks will return the interactions that have MECP2 as a participant and then the associated upstream source of the interaction or the associated downstream target of MECP2 and can be found in Table 3.8. Figure 3.4 shows examples of the directed nature of influences by MECP2. The query identified ten gene products that are known to influence or be influenced by MECP2. Three gene products were upstream of MECP2 and have an influence on MECP2, while the other 7 gene products were downstream of MECP2 and indicate that they are influenced by MECP2. This basically captures the semantics of the biological meaning of the pathway, a rare disease caused by a damaged gene that has a variety of effects and interactions.

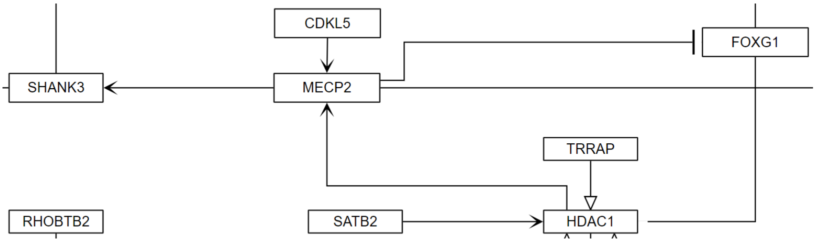


Figure 3.4: Example of direct interactions of gene products that both influence MECP2 and are influenced by MECP2 from Rett syndrome causing genes (wikipathways:WP4312). In this example, MECP2 is being influenced by HDAC1 and CDKL5. MECP2 then in turns influences SHANK3 and inhibits the activity of FOXG1.

Sphingolipid metabolism

For sphingolipid metabolism, a Python script was devised that queries the WPRDF for WikiPathways pathway wikipathways:WP1423, Ganglio Sphingolipid Metabolism, and returns a table with directed interactions that have an enzyme that is catalyzing the reaction. The query limits results to wikipathways:WP1423 as a matching criteria, then finds interactions that are annotated as being a catalysis reaction. It retrieves the associated protein for the catalysis along with the interaction that is being acted upon. Finally, the query also retrieves the source (substrate) and target (product) for the directed interaction that was being catalyzed. Figure 3.5 shows an example enzymatic reason. The results of the query are shown in Table 3.9, five conversion annotated interactions in this pathway were returned.

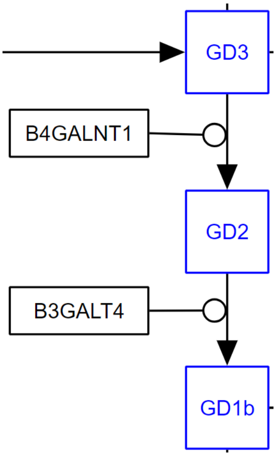


Figure 3.5: Representation of conversion of different sphingholipids to their products and the relevant enzyme catalyzing the reaction from the Ganglio Sphingolipid Metabolism pathways (wikipathways:WP1423). In this case, GD3 is converted to GD2 by the enzyme B4GALNT1. GD2 is then in turn converted to GD1b and catalyzed by B3GALT4.

Table 3.7: **Curation query showing Interaction, GPML Graph Ref from the WikiPathways RDF, and label for node at end of interaction.**

GPML Interaction	GPML Graph Ref	Participant Label
WP107_r105846/Interaction/d2818	e82	EIF4E
WP107_r105846/Interaction/cc170	ceb	ITGB4BP
WP107_r105846/Interaction/f3bb6	fc8	EIF5A
WP1403_r106688/Interaction/ide379f87c	b9666	GLUT4
WP1403_r106688/Interaction/b1235	f344c	Calcium
WP1403_r106688/Interaction/c4810	c9726	FA Synthase
WP1403_r106688/Interaction/f8d22	d9cf5	cAMP
WP1403_r106688/Interaction/d8a35	a84ee	Leptin
WP1403_r106688/Interaction/b166c	ad4a4	Malonyl-CoA
WP1403_r106688/Interaction/e0f9b	d4875	Fatty Acid Oxidation
WP1403_r106688/Interaction/af18d	dcd84	MEF2B
WP1403_r106688/Interaction/e4288	b35fe	Torc2
WP1403_r106688/Interaction/c0527	aeb8f	HMG CoA Reductase
WP1403_r106688/Interaction/cff59	d8c91	HuR
WP1403_r106688/Interaction/ae70c	b3840	Metformin
WP1403_r106688/Interaction/d14e4	b2489	Glucose
WP1403_r106688/Interaction/bedc0	af2e8	Raptor
WP1403_r106688/Interaction/c7163	f156e	PI3K (III)
WP1403_r106688/Interaction/a04e2	df1d0	HNF4A
WP1403_r106688/Interaction/d7df8	f3d7e	4E-BP1

3.4 Discussion

The analysis in this paper only involves human pathways on WikiPathways from the original, non-Reactome, collection. For other species, the results would have been affected by the more limited curation effort that has been spent on those in general. To allow us to do meaningful interaction analysis we need to have sufficient information about the interactions and their participants. Generally, a data node might be found in the graphical portion of the RDF and not in the semantic portion because of incorrect annotations, because the curator really meant to add something atypical, like an organ, or because of a failure by the conversion scripts to successfully convert the graphical information into semantic information.

Interaction types were harmonized by the scripts to turn pathway graphical information into semantic data if there was an appropriate analogue and drawing for the different notation types. This allows for example, a user to draw either a SBGN, a MIM, or a general WikiPathways inhibition drawing to have a harmonized interaction type called wp:Inhibition. In this example, since all three different notation types have the same biological meaning of indicating an inhibition event, it allows the user the flexibility to draw the pathway in the notation they are most comfortable using and still preserving the meaning of the interaction edge.

In addition to harmonization of the WikiPathways interaction types, it is shown to be possible to represent the interactions as shape expressions or ShEx. ShEx were created for all the harmonized interaction types that are found in WikiPathways. The ShEx for an interaction informs the user what is expected to be found from a certain resource. For interactions, this means it is possible to know the general shape to expect for any interaction found within the WP RDF.

For the more curated human pathways, we find that gene products that are in the GPMLRDF but not in the WPRDF, typically these are nodes that do not have a selected database resource type, like Ensembl or NCBI Gene. From Table 3.2 we learn that most of the data nodes already do have enough information to be included in the semantic part of the RDF. Future curation tasks to identify appropriate sources for the data nodes with missing annotations would enable them to become part of the semantic information. Curation efforts are a part of improving the quality of WikiPathways as a resource but also improving the coverage of interacting elements that are queryable by biologists that are looking to explore their genes or processes of interest.

Three further examples of existing problems with data nodes exist for nodes of unknown type, pathway nodes, and complex data nodes. The unknown nodes do not have an associated data type or an associated database. Pathways nodes are currently part of the WPRDF data model, but only typed as data node and not as pathway, and therefore only get counted as data node.

In the case of data nodes for complexes, there were only 18 complex nodes that do not have an equivalent in the semantic information. These complex data nodes also share the problem of missing database resource or missing data node identifiers, and therefore cannot be converted into WPRDF.

We also saw how data node types and interaction types complement each other. For example, Table 3.4 shows specific interactions as well as the type of the interaction and the interaction's participants. This can also be a useful aid in helping to identify areas of curation that need to be addressed. For example, if the participants retrieved for a conversion reaction are metabolites then this makes sense, but if the participants are proteins then there is a possibility that a post-translational modification is described but it is also possible that the user used the wrong annotation for the interaction type, especially when the two proteins are known to be derived from different genes. Based upon the results summarized in Table 3.5, we can get an estimate of what combinations of participant and interactions types are most prevalent. This gives us an indication of the accuracy of the data. For example, we found a large number of directed interactions connect two metabolites without a specific type. These are likely conversions but they still miss that typing.

We further found that one reason why interactions are captured by the GPMLRDF but not the WPRDF is because some interactions are lines connecting one or more text labels. These are not converted into the semantic layer. The WikiPathways database also allows information added as graphical annotation for the user to better understand a pathway diagram and to provide background information. This type of graphical annotation is only visually curated data but is not meant to show up in the WPRDF.

A third reason why some interactions are not captured in the semantic layer is because one of participants is a user defined group or complex. Ideally, when the participant really is a complex, then that complex itself should be identified with an external identifier like one from the Complex Portal at EBI (https://www.ebi.ac.uk/complexportal/home) (*4*). In that case it is clear that all elements of such a complex are involved in the reaction, although the curator may still have made clear that one element is directly involved. In that case, the interaction will be graphically connected with an element inside the complex.

Also in the GitHub repository is a directory titled *pastReleases* with tables of values for the queries that were performed on the November 2016 release of the WikiPathways RDF as a comparison to the June 2019 release used in this paper. The S4 file is also included as a zip file for the results of the June 2016 release. What is reflected in this comparison is that there is ongoing growth of the WikiPathways database and its semantic descriptions which sees a 43.8% increase in datanodes and an 23.3% increase in interactions from the 2016 release to the more recent release. All datanode types and interactions saw an increase in the later release compared to the earlier release, except for the case of stimulation interactions. This value went down between the releases as a result of curation efforts that identified that several of the interactions annotated as stimulations were incorrectly typed as such. Because of this curation the interactions were re-typed as their appropriate interaction type and thus we see a decrease in their number of interactions.

There is an ongoing discussion on user defined groups too, e.g. on how those should be connected and represented in the RDF as there might not be a single solution to address all the use cases of user groups. For example, these user groups often represent a class of enzymes that are all capable of catalyzing the same reaction, this can be seen

Table 3.8: **MECP2 Upstream and downstream targets.** In the table a source node is shown with its label, as well as the target and its label. The pathway in which the interaction is found and the interaction id are also provided.

Source	Source Label	Target	Target Label	Pathway	Interaction
ensembl:ENSG00000169057	MECP2	chebi:CHEBI:29987	glutamate	wikipathways:WP3584_r96364	Interaction/id4f207df3
ensembl:ENSG00000169057	MECP2	chebi:CHEBI:29987	Glutamate	wikipathways:WP3584_r96364	Interaction/id4f207df3
ensembl:ENSG00000169057	MECP2	ensembl:ENSG00000118260	CREB1	wikipathways:WP3584_r96364	Interaction/ida4a8b443
ensembl:ENSG00000169057	MECP2	ensembl:ENSG00000118260	CREB	wikipathways:WP3584_r96364	Interaction/ida4a8b443
ensembl:ENSG00000169057	MECP2	ensembl:ENSG00000176697	BDNF	wikipathways:WP3584_r96364	Interaction/id4a259c62
ensembl:ENSG00000169057	MECP2	ensembl:ENSG00000155511	GRIA1	wikipathways:WP3584_r96364	Interaction/id3bcd32
ensembl:ENSG00000169057	MECP2	ensembl:ENSG00000155511	AMPA	wikipathways:WP3584_r96364	Interaction/id3bcd32
ensembl:ENSG00000169813	HNRNPF	ensembl:ENSG00000169057	MECP2	wikipathways:WP3584_r96364	Interaction/id1c3def3d
ensembl:ENSG00000196132	MYT1	ensembl:ENSG00000169057	MECP2	wikipathways:WP3584_r96364	Interaction/id8e7af5c
ensembl:ENSG00000169045	HNRNPH1	ensembl:ENSG00000169057	MECP2	wikipathways:WP3584_r96364	Interaction/ida6a9fa9d

Table 3.9: **Sphingolipid Conversion Interactions.** In the table the enzyme for the conversion is given along with the metabolite source and its label along with the metabolite target along with its label and completed with the interaction id for the conversion

Enzyme	Metabolite Source	Source Label	Metabolite Target	Metabolite Target Label	Interaction
ensembl:ENSG00000115525	hmdb:HMDB0006750	Lactosylceramide	hmdb:HMDB0004844	GM3	Interaction/idb121743e
ensembl:ENSG00000115525	hmdb:HMDB0006750	LacCer	hmdb:HMDB0004844	GM3	Interaction/idb121743e
ensembl:ENSG00000169359	hmdb:HMDB0004913	GD3	pubchem.compound:73427362	O-Acetylated GD3	Interaction/id5f3f21f
ensembl:ENSG00000235863	hmdb:HMDB0004925	GD2	hmdb:HMDB0004926	GD1b	Interaction/idde73da53
ensembl:ENSG00000101638	hmdb:HMDB0004927	GT1b	hmdb:HMDB0004928	GQ1bA	Interaction/idc09b2721

in the example of the sphingolipid metabolism pathway, wikipathways:WP1423. Several intended interactions are not included in the WPRDF since the participants belong to a group of isoenzymes and will not be found in SPARQL query results. For this case, a simple solution would be to connect each element of the group via a duplicate interaction that is annotated as a catalysis towards the conversion, but not connect the isoenzymes to each other as is implied in the case of a biological complex. However, a user group could currently be any sort of convenient grouping and so this solution would not be a catch all solution for all groups, and further specifications would have to be included in the WikiPathways drawing options set itself.

The modelled biological knowledge of WikiPathways has previously been reported in the Waagmeester *et al.* paper (*11*). During that analysis, the first release of WPRDF was explored to determine how elements were connected to one another in that semantic part of the RDF. As discussed above, there were many interactions that are drawn in the pathway and in the graphical information about a pathway but not found in the semantic layer. This was partly addressed by curation efforts that made sure that data nodes are drawn, typed and identified correctly and interactions are drawn for instance from anchors of the data nodes to another anchor in the drawing program. Overall 56% of interactions in the graphical information is now represented in the semantic portion. The WikiPathways connection information helps the WikiPathways team with their curation efforts with automated queries that have been implemented on the Jenkins platform (*32*).

Nevertheless, as was shown in the two biological examples above, it is possible to take advantage of the semantic information in the RDF to answer relevant questions. MECP2 was chosen as it is a signaling pathway and ganglio sphingolipid metabolism is a metabolic pathway. Both MECP2 and spingolipids are active research lines in the group. For MECP2, known to be a core epigenetic regulator, it was possible to identify MECP2 in pathway diagrams and then use connectivity information to find which other elements have a direct influence upon it and which elements MECP2 influences directly. In sphingolipid metabolism, conversion of metabolites from one form to another by a catalysis reaction were shown. This has interesting implications as it is then possible to expand this knowledge to infer information about the hierarchy of enzymes in this pathway. Meaning that, for example, GD3 is converted to GD2 by enzyme B4GALNT1 and GD2 is converted to GD1B by enzyme B3GALT4. This means that anything that acts upon and affects the activity of the upstream B4GALNT1 enzyme, will also affect the conversion of GD2 to GD1B by B3GALT4 through influence on substrate availability. This is more of an indirect influence of one element upon another but it is possible to then retrieve these indirect interactions.

The connectivity information from WikiPathways has already been deployed and taken advantage of in several instances. Pathway connectivity RDF information was integrated into the Open PHACTS Discovery Platform (*33*). The connectivity information used in Open PHACTS was necessary to answer basic competency questions for the platform (*34*). The connectivity information also became a useful way to create a network of pathways to identify active subnetworks in rare diseases (*7*). This is part of a larger process involved with creating RDF of pathway data and using that information to answer questions in biology.

Conclusion

It was demonstrated that most of the graphical biological knowledge from WikiPathways is modelled in the semantic layer (WPRDF) of WikiPathways RDF with the semantic information intact and connectivity information preserved. This semantic translation allows us to answer biological questions. The MECP2 example shows directional regulatory information captured by the WPRDF, and for the other example of sphingolipid metabolism complex successive biochemical reactions are captured. MECP2 involvement in regulatory, epigenetic interactions has implications for the understanding of the rare disease Rett syndrome. Sphingolipids are important parts of cell function and structure. Being able to evaluate the order in which biological elements affect each other allows, for example, the identification of up or downstream targets that will have a similar effect when modified.

The usability of the WikiPathways pathway and connectivity information has shown to be useful and has been integrated into platforms such as the Open PHACTS Drug Discovery Platform (*33*). Improvements in WikiPathways curation and in the conversion to WikiPathways RDF support these other platforms and will allow giving a more complete picture of connectivity in biological systems. Continued curation efforts will incrementally improve many of the shortcomings of data and will continually make the semantic information better. The addition of shape expressions is a new method introduced that allows researchers to identify the form to expect from an interaction. Efforts to improve on the conversion scripts can address lost connectivity information that is for instance the result of using groups and complexes. Pathways themselves are also continually being added to WikiPathways and will continue to add to the richness of knowledge of biological interactions.

List of Abbreviations

RDF - Resource Description Framework, GPML - Graphical Pathway Markup Language, GPMLRDF - RDF for Graphical Pathway Markup Language, MIM - Molecular Interaction Map, SBGN - Systems Biology Graphical Notation, WP - WikiPathways, WikiPathways RDF - The combination of GPMLRDF and WPRDF, WPRDF - RDF for WikiPathways, SPARQL - SPARQL Protocol and RDF Query Language, KEGG - Kyoto Encyclopedia of Genes and Genomes, HGNC - HUGO Gene Nomenclature Committee, ShEx - Shape Expressions

Supplemental

Supplemental material and files can be found in the original article. The DOI of which is https://doi.org/10.1371/journal.pone.0263057.

References

1. R. A. Miller *et al.* Understanding signaling and metabolic paths using semantified and harmonized information about biological interactions. *PLOS ONE*. 2022. **17** (4): e0263057.
2. A. Piovesan *et al.* Human protein-coding genes and gene feature statistics in 2019. *BMC Research Notes*. 2019. **12** (1): 315.
3. D. S. Wishart *et al.* HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Research*. 2018. **46** (D1): D608–D617.
4. B. H. M. Meldal *et al.* Complex Portal 2018: extended content and enhanced visualization tools for macromolecular complexes. *Nucleic Acids Research*. 2018. **47** (D1): D550–D558.
5. P. C. Havugimana *et al.* A Census of Human Soluble Protein Complexes. *Cell*. 2012. **150** (5): 1068–1081.
6. S. Kikugawa *et al.* PCDq: human protein complex database with quality index which summarizes different levels of evidences of protein complexes predicted from H-Invitational protein-protein interactions integrative dataset. *BMC Systems Biology*. 2012. **6** (Suppl 2): S7.
7. R. A. Miller *et al.* Beyond Pathway Analysis: Identification of Active Subnetworks in Rett Syndrome. *Frontiers in Genetics*. 2019. **10**: 59.
8. P. D. Karp, P. E. Midford, R. Caspi, A. Khodursky. Pathway size matters: the influence of pathway granularity on over-representation (enrichment analysis) statistics. *BMC Genomics*. 2021. **22** (1): 191.

9. M. Kutmon *et al.* WikiPathways: capturing the full diversity of pathway knowledge. *Nucleic Acids Research*. 2016. **44** (D1): D488–D494.

10. M. Kutmon *et al.* PathVisio 3: An Extendable Pathway Analysis Toolbox. *PLOS Computational Biology*. 2015. **11** (2): 1–13.

11. A. Waagmeester *et al.* Using the Semantic Web for Rapid Integration of WikiPathways with Other Biological Online Data Resources. *PLOS Computational Biology*. 2016. **12** (6): e1004989.

12. A. Luna *et al.* A formal MIM specification and tools for the common exchange of MIM diagrams: an XML-Based format, an API, and a validation method. *BMC Bioinformatics*. 2011. **12** (1): 167.

13. N. L. Novère *et al.* The Systems Biology Graphical Notation. *Nature Biotechnology*. 2009. **27** (8): 735–741.

14. S. Staworko *et al.* Complexity and Expressiveness of ShEx for RDF. *Leibniz International Proceedings in Informatics*. 2015. : .

15. E. Prud'hommeaux, J. Labra Gayo, H. Solbrig. Shape expressions: an RDF validation and transformation language. *SEM '14: Proceedings of the 10th International Conference on Semantic Systems Pages*. 2014. : 32–40.

16. K. Thornton *et al.* "Using Shape Expressions (ShEx) to Share RDF Data Models and to Guide Curation with Rigorous Validation". Paper presented at: The Semantic Web; ; Cham: Springer International Publishing; 2019. Pp. 606–620.

17. A. Waagmeester *et al.* A protocol for adding knowledge to Wikidata: aligning resources on human coronaviruses. *BMC Biology*. 2021. **19** (1): 12.

18. J. D. Lewis *et al.* Purification, sequence, and cellular localization of a novel chromosomal protein that binds to methylated DNA. *Cell*. 1992. **69** (6): 905–914.

19. M. Wan *et al.* Rett Syndrome and Beyond: Recurrent Spontaneous and Familial MECP2 Mutations at CpG Hotspots. *The American Journal of Human Genetics*. 1999. **65** (6): 1520–1529.

20. U. Moog *et al.* Neurodevelopmental disorders in males related to the gene causing Rett syndrome in females (MECP2). *European journal of paediatric neurology: EJPN: official journal of the European Paediatric Neurology Society*. 2003. **7** (1): 5–12.

21. J. L. Neul *et al.* Specific mutations in methyl-CpG-binding protein 2 confer different severity in Rett syndrome. *Neurology*. 2008. **70** (16): 1313–1321.

22. F. Ehrhart *et al.* Integrated analysis of human transcriptome data for Rett syndrome finds a network of involved genes. *The World Journal of Biological Psychiatry*. 2020. **21** (10): 712–725.

23. C. R. Gault, L. M. Obeid, Y. A. Hannun. An overview of sphingolipid metabolism: from synthesis to breakdown. *Advances in experimental medicine and biology*. 2010. **688**: 1–23.

24. R. Tidhar, A. H. Futerman. The complexity of sphingolipid biosynthesis in the endoplasmic reticulum. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*. 2013. **1833** (11): 2511–2518.

25. A. H. Merrill. *De Novo* Sphingolipid Biosynthesis: A Necessary, but Dangerous, Pathway. *Journal of Biological Chemistry*. 2002. **277** (29): 25843–25846.

26. T. Kolter, K. Sandhoff. Sphingolipid metabolism diseases. *Biochimica et Biophysica Acta (BBA) - Biomembranes*. 2006. **1758** (12): 2057–2079.

27. R. Kraut, N. Bag, T. Wohland, en, *Methods in Cell Biology*, Elsevier, Amsterdam, 2012, vol. 108, chap. 18, pp. 395–427, ISBN: 978-0-12-386487-1, (2019; https://linkinghub.elsevier.com/retrieve/pii/B9780123864871000183).

28. A. Bohler *et al.* Reactome from a WikiPathways Perspective. *PLOS Computational Biology*. 2016. **12** (5): e1004941+.

29. A. Fabregat *et al.* The Reactome pathway Knowledgebase. *Nucleic acids research*. 2016. **44**: D481–7.

30. K. Thomas *et al.* Jupyter Notebooks - a publishing format for reproducible computational workflows. *Stand Alone*. 2016. : .

31. M. Hucka *et al.* The Systems Biology Markup Language (SBML): Language Specification for Level 3 Version 2 Core Release 2. *Journal of Integrative Bioinformatics*. 2019. **16** (2): .

32. D. N. Slenter *et al.* WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Research*. 2017. **46** (D1): D661–D667.

33. R. Miller *et al.* Explicit interaction information from WikiPathways in RDF facilitates drug discovery in the Open PHACTS Discovery Platform [version 2; peer review: 2 approved]. *F1000Research*. 2018. **7** (75): .

34. K. Azzaoui *et al.* Scientific competency questions as the basis for semantically enriched open pharmacological space development. *Drug Discovery Today*. 2013. **18** (17-18): 843–852.

4

Explicit interaction information from WikiPathways RDF in the Open PHACTS Discovery Platform

Adapted from: R. A. Miller *et al.* Explicit interaction information from WikiPathways in RDF facilitates drug discovery in the Open PHACTS Discovery Platform. *F1000Research*. 2018. **7**: 75.

Abstract

Open PHACTS is a pre-competitive project to answer scientific questions developed recently by the pharmaceutical industry. Having high quality biological interaction information in the Open PHACTS Discovery Platform is needed to answer multiple pathway related questions. To address this, updated WikiPathways data has been added to the platform. This data includes information about biological interactions, such as stimulation and inhibition. The platform's Application Programming Interface (API) was extended with appropriate calls to reference these interactions. These new methods of the Open PHACTS API are available now.

4.1 Introduction

Targeting proteins to ideally restore normal biological processes is a common starting point in drug discovery (*2*). The Open PHACTS Discovery Platform (OPDP) was designed to help identify protein targets and information about their associations with each other (*3*–*5*). The OPDP supports target identification and validation by including target-target interactions from WikiPathways (*6*–*8*). Of these interaction networks, proteins sharing a downstream path allows investigation of alternative drug target combinations. Even the knowledge of which biological pathways participate in disease-related processes provides insight in the pathway topology between the targets. The importance and need of providing access to interaction information for real-world research questions was outlined in a recent Open PHACTS paper (*9*).

The Open PHACTS project was born out of the desire to integrate pharmacological data from multiple precompetitive sources to efficiently address scientific questions that cannot be answered with single data sources (*9*). It integrates data using linked data approaches (*4*) from chemical and biological sources such as ChEBI, ChEMBL, UniProt, and WikiPathways (*7*). However, the OPDP did not previously include calls to access specific up- and downstream interaction effects. This information is needed for questions related to drug repositioning and repurposing. Up- or downstream targets may be interesting alternatives with similar therapeutic effect to targets, for which it is particularly hard to develop a drug agent. Thus, finding a target that has already been drugged or is more drug tractable will be advantageous. Here we describe how to identify alternative targets in the same cellular pathway using OPDP against the WikiPathways data.

4.2 Methods

4.2.1 Implementation

The WikiPathways Resource Description Framework data (WPRDF) is released as part of the monthly releases (*6*). The native format for WikiPathways is Graphical Pathway Markup Language (GPML) based on the eXtensible Markup Language (XML) standard. The RDF export is transformed from the original GPML. In the RDF representation we use two distinct controlled vocabularies, to distinguish between the graphical notation of a pathway and the biological meanings expressed in the pathway. This is done to allow integration with other pathway repositories which use other graphical notations or none. The WikiPathways RDF also includes details about directed and undirected interactions. Directed biochemical interactions capture the source and target which are depicted as an arrow in simple pathway drawings. WikiPathways adds biological meaning to interactions with Molecular Interaction Map (MIM) interaction types, like inhibitions, enzyme catalyzed reactions, and stimulations (*10*), as well as Systems Biology Graphical Notation (SBGN) interactions (*11*). Reactome pathways in WikiPathways use SBGN interactions (*12, 13*). However, because MIM and SBGN use different drawing styles, we normalize their inhibition types into a common inhibition type, defined by the WikiPathways ontology (https://vocabularies.wikipathways.org/wp).

The WikiPathways basic drawing tools also contain generic arrows and T-bar annotations that give the user the ability to create basic diagrams without the semantic meaning of MIM or SBGN notations. The interactions connecting these nodes are captured, but the only explicit information is that it is a directed interaction from a source to a target. To handle more complicated enzyme reaction drawings, where there is not a single line that directly connects targets in a cascade of enzymatic reactions, a query was developed that recognizes these types of reactions. However, this is not implemented in the current Open PHACTS Application Programming Interface (API).

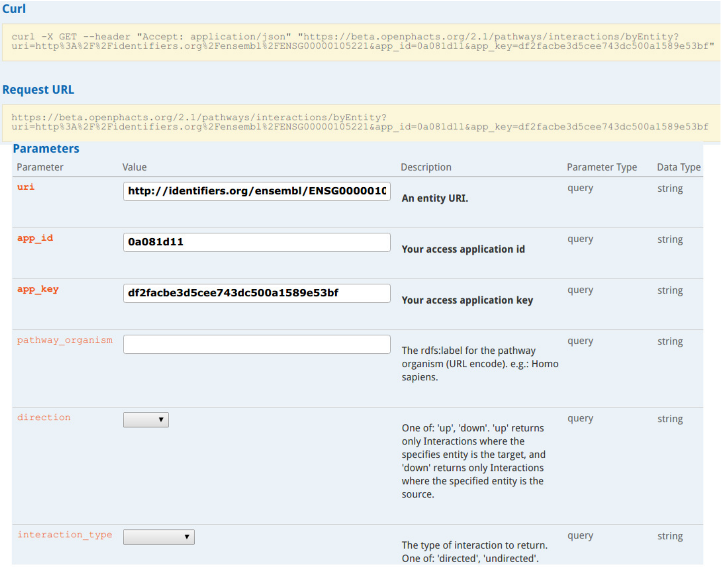
Version 2.1 of the OPDP API contains three new calls for interactions and their pathways. The first call, */pathway/getInteractions*, returns all interactions involved in a pathway. To use this feature, the user specifies a pathway URI and OPDP returns its interactions including information about direction and the connected entities. The direction information is relayed as a starting node having a *wp:source* annotation, while the end of the interaction has the *wp:target* annotation. In its simplest form, this means that if gene product A is interacting with a gene product B, then we have *wp:source* for product A and *wp:target* for product B. However, the presented new methods also support interactions with multiple sources and targets for more complex interactions that are more accurately represented this way.

The second added call, */pathways/interactions/byEntity*, returns the direction of the interactions involving this entity. An entity is specified by a URI and can be a metabolite, protein, gene product, or RNA. API options allow the user to select only upstream or only downstream interactions. If a direction is not specified in the call, all the adjacent interactions will be retrieved regardless of their direction. The results also specify the interaction type (e.g. inhibition, stimulation, conversion). Vocabularies.wikipathways.org also identifies catalysis and binding events as well as a more generic directedInteraction in the case where the type of the interaction is not identified. This ability to select the interaction direction is specifically what allows users to answer scientific questions around upstream and downstream effects, such as those defined by Open PHACTS. The third API call is */pathways/interactions/byEntity/count* which is a helper function that returns the number of interactions for a target.

4.2.2 Operation

The OPDP API calls are backed by SPARQL searches against the loaded WikiPathways RDF. The query parameters that are required or optional are given in the documentation of Open PHACTS (https://dev.openphacts.org/docs/2.1). As in previous versions, the API uses HTTP

Example input for */pathways/interactions/byEntity* call for AKT2



Curl

```
curl -X GET --header "Accept: application/json" "https://beta.openphacts.org/2.1/pathways/interactions/byEntity?
uri=http%3A%2F%2Fidentifiers.org%2Fensembl%2FENSG00000105221&app_id=0a081d11&app_key=df2facbe3d5cee743dc500a1589e53bf"
```

Request URL

```
https://beta.openphacts.org/2.1/pathways/interactions/byEntity?
uri=http%3A%2F%2Fidentifiers.org%2Fensembl%2FENSG00000105221&app_id=0a081d11&app_key=df2facbe3d5cee743dc500a1589e53bf
```

Parameters

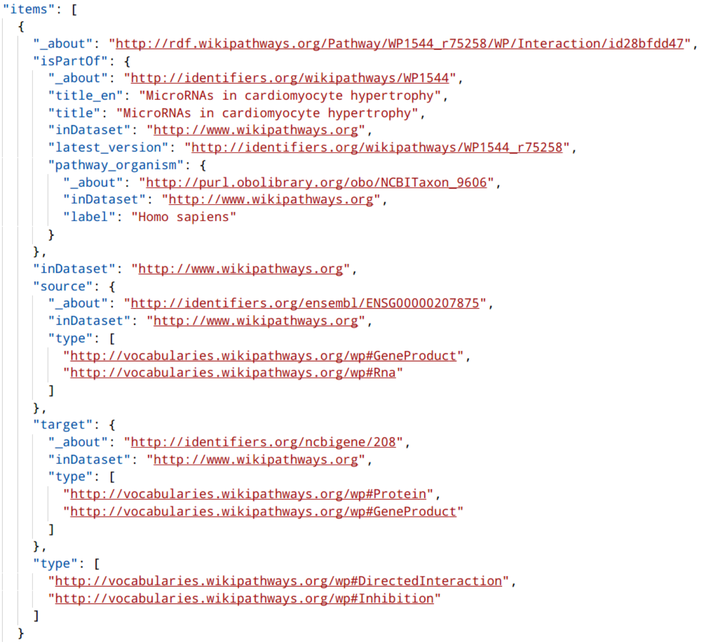
Parameter	Value	Description	Parameter Type	Data Type
uri	http://identifiers.org/ensembl/ENSG0000010	**An entity URI.**	query	string
app_id	0a081d11	**Your access application id**	query	string
app_key	df2facbe3d5cee743dc500a1589e53bf	**Your access application key**	query	string
pathway_organism		The rdfs:label for the pathway organism (URL encode). e.g.: Homo sapiens.	query	string
direction	▾	One of: 'up', 'down'. 'up' returns only Interactions where the specifies entity is the target, and 'down' returns only Interactions where the specified entity is the source.	query	string
interaction_type	▾	The type of interaction to return. One of: 'directed', 'undirected'.	query	string

Figure 4.1: Parameters (bottom) and *curl* command (top) for the GET */pathways/interactions/byEntity* call. The GET portion tells the API to retrieve data with the associated call. It takes an entity URI, the Ensembl ID for AKT2, and returns a list interactions for AKT2. The obligatory parameters are shown in bold. Entity IDs that are acceptable for queries include Ensembl, Entrez Gene, and UniProt for genes, proteins, and RNAs. For metabolites the ID sources HMDB, ChEBI, and ChemSpider, for example, are acceptable entity IDs

GET to call methods and needs a (free) application ID and key (see https://dev.openphacts.org/signup) (*4*).

To ensure multiple URI schemes can be used to identify genes, proteins, and metabolites, the Open PHACTS platform uses an Identifier Mapping Service (IMS) (*7*). This ensures that people can use Ensembl, NCBI Gene, and others for genes, UniProt, Ensembl, etc. for proteins, and HMDB, ChEBI, CAS registry number, and PubChem for metabolites. Furthermore, it supports identifiers.org formatted URIs, further simplifying entering identifiers (*14*).

Example query results for *lpathwayslinteractionslbyEntity* call for AKT2



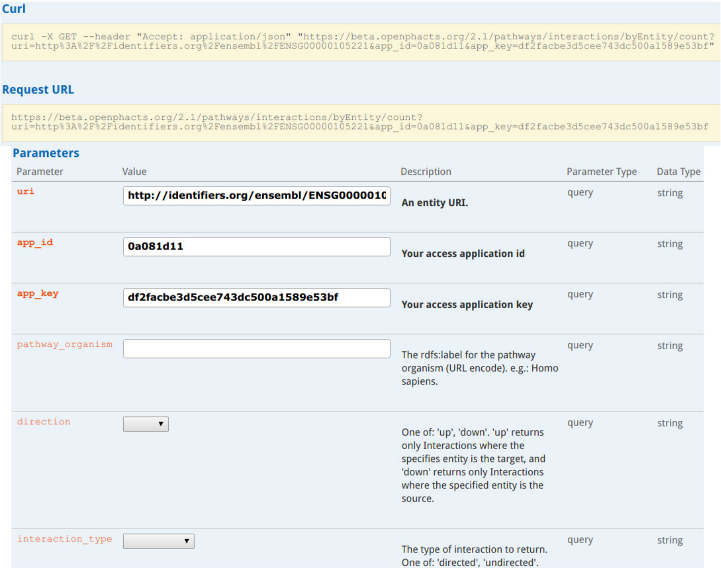
```
"items": [
  {
    "_about": "http://rdf.wikipathways.org/Pathway/WP1544_r75258/WP/Interaction/id28bfdd47",
    "isPartOf": {
      "_about": "http://identifiers.org/wikipathways/WP1544",
      "title_en": "MicroRNAs in cardiomyocyte hypertrophy",
      "title": "MicroRNAs in cardiomyocyte hypertrophy",
      "inDataset": "http://www.wikipathways.org",
      "latest_version": "http://identifiers.org/wikipathways/WP1544_r75258",
      "pathway_organism": {
        "_about": "http://purl.obolibrary.org/obo/NCBITaxon_9606",
        "inDataset": "http://www.wikipathways.org",
        "label": "Homo sapiens"
      }
    },
    "inDataset": "http://www.wikipathways.org",
    "source": {
      "_about": "http://identifiers.org/ensembl/ENSG00000207875",
      "inDataset": "http://www.wikipathways.org",
      "type": [
        "http://vocabularies.wikipathways.org/wp#GeneProduct",
        "http://vocabularies.wikipathways.org/wp#Rna"
      ]
    },
    "target": {
      "_about": "http://identifiers.org/ncbigene/208",
      "inDataset": "http://www.wikipathways.org",
      "type": [
        "http://vocabularies.wikipathways.org/wp#Protein",
        "http://vocabularies.wikipathways.org/wp#GeneProduct"
      ]
    },
    "type": [
      "http://vocabularies.wikipathways.org/wp#DirectedInteraction",
      "http://vocabularies.wikipathways.org/wp#Inhibition"
    ]
  }
```

Figure 4.2: Result in the JSON format of the AKT2 query from Figure 4.1. The participants of the interaction are directed from source (hsa-let7b) to target (AKT2). It also shows the type of interaction (inhibition), and the biological types of the interaction participants.

4.3 Example Queries

We are demonstrating the platform with three example calls. All the API calls require use of an application ID and an application key. This key and ID can be acquired by creating a free Open PHACTS account. The first example is an application to the PI3K/AKT pathway for cell growth regulation which contain important targets for cancer treatment (*15*) 14. The AKT protein has a central role and usefully shows the API call's ability to return connected elements with the */pathways/interactions/byEntity* and the */pathway/getInteractions* calls. The API calls

Example input for */pathways/interactions/byEntity/count* call for AKT2



Curl

```
curl -X GET --header "Accept: application/json" "https://beta.openphacts.org/2.1/pathways/interactions/byEntity/count?
uri=http%3A%2F%2Fidentifiers.org%2Fensembl%2FENSG00000105221&app_id=0a081d11&app_key=df2facbe3d5cee743dc500a1589e53bf"
```

Request URL

```
https://beta.openphacts.org/2.1/pathways/interactions/byEntity/count?
uri=http%3A%2F%2Fidentifiers.org%2Fensembl%2FENSG00000105221&app_id=0a081d11&app_key=df2facbe3d5cee743dc500a1589e53bf
```

Parameters

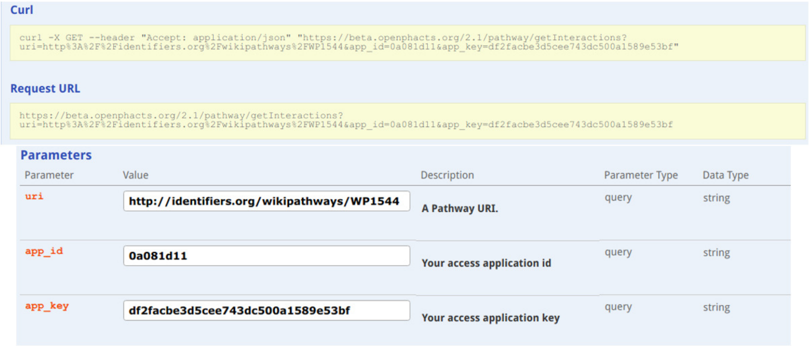
Parameter	Value	Description	Parameter Type	Data Type
uri	http://identifiers.org/ensembl/ENSG0000010	An entity URI.	query	string
app_id	0a081d11	Your access application id	query	string
app_key	df2facbe3d5cee743dc500a1589e53bf	Your access application key	query	string
pathway_organism		The rdfs:label for the pathway organism (URL encode). e.g.: Homo sapiens.	query	string
direction	▾	One of: 'up', 'down'. 'up' returns only Interactions where the specifies entity is the target, and 'down' returns only Interactions where the specified entity is the source.	query	string
interaction_type	▾	The type of interaction to return. One of: 'directed', 'undirected'.	query	string

Figure 4.3: Parameters (bottom) and *curl* command (top) for the GET */pathways/interactions/byEntity/count* call. It takes a URI for an entity, in this case the Ensembl ID for AKT2 and returns a count of the interactions to which this gene product is involved. Only the entity URI, app ID, and app key are required fields. Optional parameters are pathway organism, direction, or type of interaction.

can help aid drug discovery by taking a target, in this case AKT, and easily identify other connected proteins that could potentially be used as drug targets with a common downstream effect.

Figure 4.1 shows the web interface of the API call that returns the connectivity of the AKT2 target to both upstream or downstream proteins or gene products. This method allows the user to identify connections to other targets in the pathway. The results of that API call (Figure 4.2) show the AKT2 interaction with microRNA. A helper method (Figure 4.3): */pathways/interactions/byEntity/count* is also included. It re-

Example input for */pathways/getInteractions* call for MicroRNAs in cardiomyocyte hypertrophy pathway



Curl

```
curl -X GET --header "Accept: application/json" "https://beta.openphacts.org/2.1/pathway/getInteractions?
uri=http%3A%2F%2Fidentifiers.org%2Fwikipathways%2FWP1544&app_id=0a081d11&app_key=df2facbe3d5cee743dc500a1589e53bf"
```

Request URL

```
https://beta.openphacts.org/2.1/pathway/getInteractions?
uri=http%3A%2F%2Fidentifiers.org%2Fwikipathways%2FWP1544&app_id=0a081d11&app_key=df2facbe3d5cee743dc500a1589e53bf
```

Parameters

Parameter	Value	Description	Parameter Type	Data Type
uri	http://identifiers.org/wikipathways/WP1544	A Pathway URI.	query	string
app_id	0a081d11	Your access application id	query	string
app_key	df2facbe3d5cee743dc500a1589e53bf	Your access application key	query	string

Figure 4.4: Parameters (bottom) and *curl* command (top) for the */pathways/getInteractions* call. It is intended to take the pathway URI from WikiPathways and return a list of interaction involved in that particular pathway. Pathway URI, app ID, and app key are the only required values for this call.

turns the number of all interactions in which an entity is participates. This helps the user get a sense of the prevalence of the queried entity with interactions in pathways found on WikiPathways. An example result for this query can be found in Supplementary Figure 1.

The other call implemented, */pathway/getInteractions* (Figure 4.4), demonstrates an API call to return all interactions in the MicroRNAs in cardiomyocyte hypertrophy pathway (*16*). This pathway has interaction details for AKT, mTOR, and PI3K, which are all important targets in cancer research (*17*). For each interaction the participants are given and whether it is a directed or undirected interaction. An example result for this query can be seen in Supplementary Figure 2.

Example workflows

In order to demonstrate the basic use of the introduced API methods, we developed two workflows, available in the Supplementary Material. One uses Python to return a file with the results in a table and the other uses a HTML webpage using the ops.js JavaScript client library (*18*). More involved workflows have been developed for KNIME and Pipeline Pilot (*19*, *20*).

The Python script example uses the Open PHACTS */pathway/getInteraction* API call and prompts the user to enter a WikiPathways pathway number that they wish to query, such as 1544 for WikiPathways pathway WP1544. Invocation of the API call with the pathway identifier returns information about the directed interactions that are involved with the pathway. The information that is returned is the interaction ID used by WikiPathways, the interaction type, and URIs for the source and target of the interaction. In order to convert the URIs into something more readable, a SPARQL query is then executed to get labels, from the WikiPathways SPARQL endpoint, for the source and target of the interaction. The results are written to a file with the interaction ID, interaction type, URIs for the source and target, as well as alias IDs, the *curl* for the API call, the pathway ID used, and a number of interactions returned.

The second example uses a HTML5 webpage and the ops.js JavaScript client library to retrieve interactions for a particular gene, using the URI for the gene's Ensembl identifier and the */pathways/interactions/byEntity* API method. The ops.js library passes the returned JSON with interaction information to a callback function, where the interacting source and target are extracted and the interacting entity determined. For each interacting entity, which may be a protein, RNA, or small compound, a call to the */pathways/interactions/byEntity/count* method is made to return the number of interaction that entity has.

4.4 Summary

While the calls identified here are simple calls, workflow tools make it possible to take advantage of the integrative nature of the OPDP to make API calls in succession. Two such workflow tools that work with the OPDP are KNIME and Pipeline Pilot. With these tools, it is possible to perform a directional query of a target and identify alternative targets that can then be queried against the chemistry calls to identify active compounds for these alternative targets. The client libraries ops.js, ops4j, and ropenphacts also support Open PHACTS and the interaction calls for pathways. This allows users to perform API calls to the OPDP using their preferred language or platform, such as JavaScript, Java, or R.

The addition of interactions with direction information allows OPDP to answering more of the pre-defined scientific questions (*3*). The directional information allows the user to explore how proteins and gene products are connected with one another and easily access this information. This is illustrated in the example queries using the cancer target AKT.

Software availability

Online service: https://dev.openphacts.org/docs/2.1
Latest source code is available at:
https://github.com/openphacts/OPS_LinkedDataApi
Archived source code of discussed version:
https://doi.org/10.5281/zenodo.1068252 (*21*)

Supplemental

Supplemental material can be found in the original article. The DOI of which is https://doi.org/10.12688/f1000research.13197.2.

References

1. R. A. Miller *et al.* Explicit interaction information from WikiPathways in RDF facilitates drug discovery in the Open PHACTS Discovery Platform. *F1000Research*. 2018. **7**: 75.
2. S. L. Schreiber. Target-Oriented and Diversity-Oriented Organic Synthesis in Drug Discovery. *Science*. 2000. **287** (5460): 1964–1969.
3. K. Azzaoui *et al.* Scientific competency questions as the basis for semantically enriched open pharmacological space development. *Drug Discovery Today*. 2013. **18** (17): 843–852.
4. A. J. Williams *et al.* Open PHACTS: semantic interoperability for drug discovery. *Drug Discovery Today*. 2012. **17** (21-22): 1188–1198.
5. D. Digles *et al.* Open PHACTS computational protocols for in silico target validation of cellular phenotypic screens: knowing the knowns. *Med. Chem. Commun.* 2016. **7**: 1237–1244.
6. A. Waagmeester *et al.* Using the Semantic Web for Rapid Integration of WikiPathways with Other Biological Online Data Resources. *PLOS Computational Biology*. 2016. **12** (6): e1004989.
7. A. J. Gray *et al.* Applying linked data approaches to pharmacology: Architectural decisions and implementation. *Semantic Web*. 2014. **12**: 101–113.
8. T. Kelder *et al.* Mining Biological Pathways Using WikiPathways Web Services. *PLOS ONE*. 2009. **4** (7): 1–4.
9. C. Chichester *et al.* Drug discovery FAQs: workflows for answering multidomain drug discovery questions. *Drug Discovery Today*. 2015. **20** (4): 399–405.
10. A. Luna *et al.* A formal MIM specification and tools for the common exchange of MIM diagrams: an XML-Based format, an API, and a validation method. *BMC Bioinformatics*. 2011. **12** (1): 167.
11. N. Le Novère *et al.* The Systems Biology Graphical Notation. *Nat Biotech*. 2009. **27** (8): 735–741.

12. M. Kutmon *et al.* WikiPathways: capturing the full diversity of pathway knowledge. *Nucleic Acids Research*. 2016. **44** (D1): D488–D494.

13. D. Croft *et al.* The Reactome pathway knowledgebase. *Nucleic Acids Research*. 2014. **42** (D1): D472–D477.

14. N. Juty, N. L. Novère, H. Hermjakob, C. Laibe. Towards the Collaborative Curation of the Registry underlying identifiers.org. *Database*. 2013. **2013**: .

15. B. Vanhaesebroeck, L. Stephens, P. Hawkins. PI3K signalling: the path to discovery and understanding. *Nat Rev Mol Cell Biol*. 2012. **13** (3): 195–203.

16. M. Levels *et al.* MicroRNAs in cardiomyocyte hypertrophy (Homo sapiens). 2017. : .

17. H. Li, J. Zeng, K. Shen. PI3K/AKT/mTOR signaling pathway as a therapeutic target for ovarian cancer. *Archives of Gynecology and Obstetrics*. 2014. **290** (6): 1067–1078.

18. I. Dunlop *et al.*, *Openphacts/Ops.Js: Ops.Js 7.0.0 For Open Phacts 2.1 Api*, 2016, (https://zenodo.org/record/167595).

19. M. R. Berthold *et al.* "KNIME: The Konstanz Information Miner". Paper presented at: Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007); ; : Springer; 2007. .

20. J. Ratnam *et al.* The Application of the Open Pharmacological Concepts Triple Store (Open PHACTS) to Support Drug Discovery Research. *PLoS ONE*. 2014. **9** (12): e115460.

21. Fundatureanu-Sever *et al.*, *Openphacts/Ops_Linkeddataapi: Open Phacts Linked Data Api 2.1.0*, 2016, (https://zenodo.org/record/1068252).

5

Beyond pathway analysis: Identification of active subnetworks in Rett syndrome

Adapted from: R. A. Miller *et al.* Beyond Pathway Analysis: Identification of Active Subnetworks in Rett Syndrome. *Frontiers in Genetics*. 2019. **10**: 59.

Abstract

Pathway and network approaches are valuable tools in analysis and interpretation of large complex omics data. Even in the field of rare diseases, like Rett syndrome, omics data are often available, and the maximum use of such data requires sophisticated tools for comprehensive analysis and visualization of the results. Pathway analysis with differential gene expression data has proven to be extremely successful in identifying affected processes in disease conditions. In this type of analysis, pathways from different databases like WikiPathways and Reactome are used as separate, independent entities. Here, we show for the first time how these pathway models can be used and integrated into one large network using the WikiPathways RDF containing all human WikiPathways and Reactome pathways to perform network analysis on transcriptomics data. This network was imported into the network analysis tool Cytoscape to perform active submodule analysis. Using a publicly available Rett syndrome gene expression dataset from frontal and temporal cortex, classical enrichment analysis including pathway and Gene Ontology analysis revealed mainly immune response, neuron specific and extracellular matrix processes. Our active module analysis provided a valuable extension of the analysis prominently showing the regulatory mechanism of *MECP2*, especially on DNA maintenance, cell cycle, transcription and translation. In conclusion, using pathway models for classical enrichment and more advanced network analysis enables a more comprehensive analysis of the gene expression data and provides novel results.

5.1 Introduction

In a diseased state, many molecular processes in the human body are affected and dysregulated. Performing pathway analysis on molecular data sets comparing healthy vs. diseased subjects is immensely effective in finding affected pathways and it enables researchers to study the underlying processes in detail, to reveal possible disease mechanisms. While standard enrichment methods have limitations and pathways are analysed independently with their arbitrary process boundaries (*2*), the pathway models themselves are very interesting from a network science perspective. These models contain detailed information about biological molecules and their interactions with one another, which can be visualized and analysed using network biology tools (*3*). The detailed models of these biological processes are collected in online pathway databases like WikiPathways (*4*) and Reactome (*5*). The availability of pathway models in the structured and semantic Resource Description Framework format (RDF) creates the possibility to integrate all pathway models into one large network and therefore incorporate the relations and overlap between them (*6*). By removing artificial boundaries, this will enable us to study the systemic effects of diseases, such as Rett syndrome, using network biology methods. Specifically, we can look for subnetworks, even if not present in pathways as found in pathway databases, which reflect modules of differential biological activity.

Rett syndrome (MIM:312750, (*7*)) is a rare genetic disorder, caused in most patients by a loss of function mutation in the *MECP2* gene (*8*). The accompanying MECP2 protein is multifunctional and acts as an epigenetic repressor, transcriptional repressor and transcriptional activator. MECP2 binds DNA on methylated CpG islands and is involved in several regulatory activities: attracting histone deacetylases (HDAC1), increasing packing density of DNA, repressing and in specific genes also activating gene expression, and due to its phosphorylation sites, MECP2 activity is sensitive to intracellular signalling (*9*, *10*). Due to its regulatory role, many downstream genes are affected in case of loss of function, resulting in a broad range of symptoms including moderate to severe intellectual disability, gait problems, stereotypic movements, dystonia, scoliosis, epileptic seizures, and sleep problems (*11*, *12*). In the past ten years, omics data analysis on the level of genome, transcriptome or proteome saw an increase in importance, to analyse and understand the holistic impact of MECP2, respectively, the impact of an impaired MECP2. (*13*) recently reviewed the available transcriptomics studies on Rett syndrome and came to the conclusion that the most researched impact of MECP2 dysfunction lies with dendritic connectivity and synapse maturation, mitochondrial dysfunction, and glial cell activity. Recent pathway analysis results of single and integrated studies identified changes in intracellular signalling, including EIF2 (eukaryotic translation initiation) signalling, cytoskeleton and cell metabolism including mitochondrial function (*14*, *15*) .

In this study, we aim to investigate the molecular changes in Rett syndrome patients using a network-based approach by integrating existing pathway models from WikiPathways and Reactome into one large network and identifying disease-affected submodules that show differential gene expression. We will compare the results with standard enrichment analysis methods, including pathway and Gene Ontology analysis, and expect that the identified disease modules will also contain interactions in pathways not found through those methods.

5.2 Material and Methods

Dataset

The publicly available dataset studying the transcriptome in human brain tissue of Rett syndrome patients and healthy controls from the Gene Expression Omnibus (GEO) was used (GEO:GSE75303). The original study was published by (*16*). The dataset contains transcriptome data obtained with Illumina HumanHT-12 V4.0 expression beadchips.

The samples were taken postmortem from human frontal and temporal cortex of three Rett syndrome patients (*MECP2* mutations c.378-2A>G, c.763C>T, c.451G>T) and three age-, gender- and ethnicity-matched controls.

Raw and normalized data as well as study metadata were obtained (GEO:GSE75303) and subjected to quality control, including signal distribution and sample grouping analyses, using plotting functions from ArrayAnalysis.org (*17*). No samples were excluded for further analysis. The normalized data was filtered to remove all probes with a detection p-value of 1 for all samples, indicating overall absence of expression. Thereafter, the limma package for R (version 3.30.13, (*18*)) was used to compute differential expression between Rett patients and controls for the frontal and temporal cortex samples separately. For each probe, this results in estimates of the log2 fold change and p-value significance between the patient and control groups. Probes were re-annotated with Ensembl gene identifiers based on Ensembl build 91 using the BridgeDbR package (version 1.8.0, (*19*)) with the Hs_Derby_-Ensembl_91.bridge database (*20*).

Enrichment analysis

We performed pathway analysis with PathVisio (version 3.3.0, (*21*)) and Gene Ontology (GO) analysis with GO-Elite (version 1.2, (*22*)).

For GO analysis with GO-Elite, the input gene lists for frontal and temporal cortex contained all significantly changed genes (p-value < 0.05) with an absolute fold change cutoff of 1.5. Ensembl identifiers of all measured genes in the datasets were provided as the background list. Number of permutations was set to 2,000. Pruned GO-term results (i.e. GO terms for which genes in subterms that were found to be significant were removed) were filtered based on Z-score (> 1.96), permuted p-value (< 0.05) and a minimum number of changed genes of five.

Pathway analysis was performed on a combined human pathway collection from all curated WikiPathways pathways including the Reactome pathway set (in total 903 pathways, October 2018 release). Differential gene expression was mapped to genes on the pathway diagrams using the Hs_Derby_Ensembl_91.bridge identifier mapping database. Thereafter, pathway statistics was performed on differential gene expression for temporal and frontal cortex using the following criteria to select only significantly differentially expressed genes (absolute fold change cutoff of 1.5 and p-value < 0.05):

```
(log2FC < -0.58 OR log2FC > 0.58) AND p-value < 0.05.
```

The resulting ranked pathway list was filtered based on Z-score (> 1.96), permuted p-value (< 0.05) and minimum number of changes (positive) genes of five.

Pathway-based network construction

Biological pathway models are small sub-networks describing specific biological processes. Connecting and integrating pathway models in one large network enables us to use network biology tools and approaches to study and investigate the network.

We used the WikiPathways RDF from October 2018 release (*6*) to retrieve information about all interactions in the pathway models of two major pathway databases, WikiPathways and Reactome. The SPARQL query language was used to retrieve the relevant data. The scripts to generate the constructed network are available on GitHub (https://github.com/wikipathways/wprdf2cytoscape). Interactions with at least two annotated interaction participants (gene product, metabolite, complex) are included. Gene products have unified Ensembl (*23*) identifiers, metabolites have either Wikidata (*24*), ChEBI (*25*) or HMDB identifiers (*26*), and complexes have Reactome identifiers. A list of frequently occurring small molecules (Supplementary Table 1), like H^+, H_20, ATP, etc, were removed from the network to prevent inclusion of paths with no specific biological relevance. Such small molecules tend to be artificial hub nodes simply because e.g. ATP is used/produced in a lot of metabolic reactions. As shown in Figure 5.1, each interaction

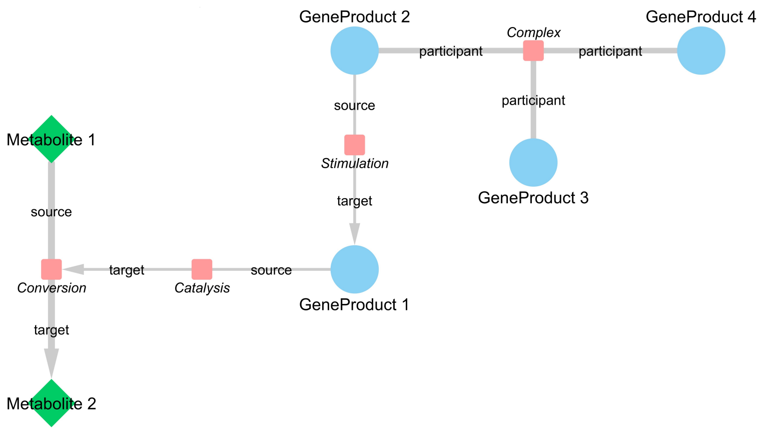


Figure 5.1: **WikiPathways network structure.** Each interaction is represented as a node in the network with links to all participants. If the interaction is directed, the information about source and target nodes is added as an edge attribute. The nodes represented as small, red rounded rectangles are interactions, blue circles represent gene products and green diamonds show metabolites. Edge thickness indicates in how many different pathways the interaction is present.

is represented by an interaction node in the network with edges to all participant nodes (either source, target or participant). For each interaction, it is recorded in which pathway or pathways the interaction is present. By connecting all the retrieved interactions, a large network representing all human pathway models was created.

Active module analysis

The constructed network was loaded into Cytoscape (version 3.7.0), a network analysis and visualization tool (*27*). Differential expression analysis data (log2 fold changes and p-values) for both frontal and

	Temporal cortex down-regulated	Temporal cortex not changed	Temporal cortex up-regulated
Frontal cortex down-regulated	88 ↓↓	44 ↓ -	1 ↓↑
Frontal cortex not changed	171 -↓	18,576 - -	55 -↑
Frontal cortex up-regulated	3 ↑↓	62 ↑-	23 ↑↑

Table 5.1: **Differentially expressed genes in frontal and temporal cortex.** 133 and 88 genes were significantly down- and up-regulated in frontal cortex, respectively. 262 and 79 genes were significantly down- and up-regulated in temporal cortex, respectively. 88 genes are down-regulated, and 23 genes are up-regulated in both brain regions. Only four genes show different expression patterns. The following filtering criteria were used: p-value < 0.05 and absolute log2 fold change > 0.58.

temporal cortex were added as node attributes to the network.

The Cytoscape app jActiveModules (version 3.2.1, (*28*)) was used to identify active submodules in the large network that show significant changes in expression. These subnetworks are freed from the artificial pathway boundaries of conventional pathway models found in WikiPathways and Reactome. The following parameters were used to find active submodules: p-value as the node attribute, number of modules was set to five, overlap threshold of 0.8, and search strategy with a search depth of two.

5.3 Results

Gene expression

The total number of probes measured was 37,707 from which 29,024 could be linked to Ensembl identifiers. After merging multiple probe identifiers for the same Ensembl identifier, 19,023 unique gene identifiers remained. Differential gene expression analysis revealed 1,953 in

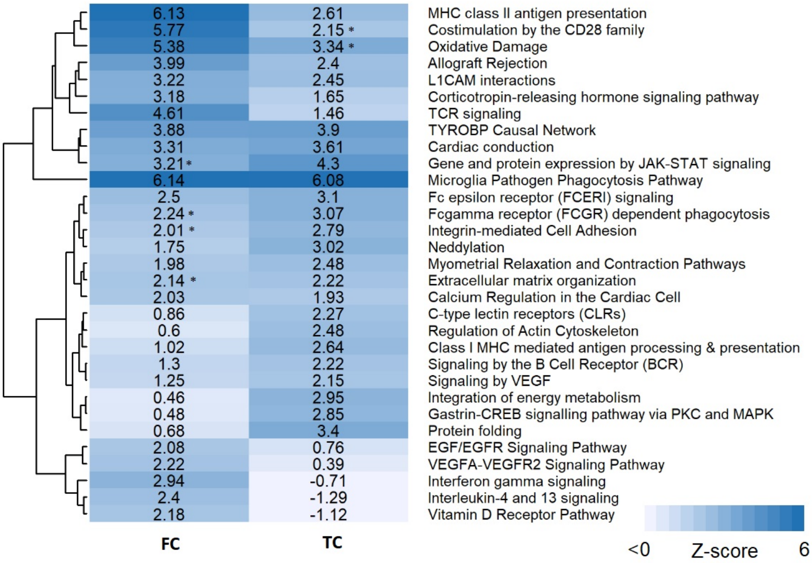


Figure 5.2: **Pathway analysis results for frontal and temporal cortex data.** Pathways are clustered in this heatmap based on their Z-scores. Pathways with a high Z-score (>1.96) contain significantly more changed genes than expected and are considered pathways of interest. An asterisk next to the Z-score value indicates pathways with a significant Z-score (>1.96) but less than five changed genes.

the frontal cortex and 2,436 significantly changed genes in the temporal cortex samples of RETT syndrome patients versus controls. Only 221 in frontal and 341 of the significantly changed genes in temporal cortex had a more than 1.5-fold increase or decrease in expression (|log2 fold change| > 0.58). In both brain regions, more genes were down-regulated in Rett syndrome patients than up-regulated, see Table 5.1, which matches the original publication (*16*).

Gene Ontology analysis

Gene Ontology overrepresentation analysis identified 39 and 50 biological processes as altered in frontal and temporal cortex, respectively (Supplementary Tables 2 and 3). Summarizing, neuron specific and immune system-related processes were found to be enriched in both brain regions for Rett syndrome patients. In temporal cortex, additionally, regulation of translational initiation (GO:0006446) and an extracellular matrix/cytoskeleton-related process (GO:0007229) were found to be enriched. Interestingly, the microglia relevant complement factors C1QB and C1QC were found in the enriched GO classes defense response (GO:0006952) and immune effector process (GO:0002252).

Pathway analysis

Pathway analysis was performed in PathVisio for both brain regions separately. Overrepresentation analysis revealed 18 and 21 pathways altered in the datasets for frontal and temporal cortex, respectively (Z-score > 1.96, minimum five changed genes), see Figure 5.2. Interestingly, eight pathways were altered in both frontal and temporal cortex. Similar to the results of the GO analysis, several immune system-related and extracellular matrix/cytoskeleton-related pathways were found to be enriched. Additionally, calcium channel related processes including muscle contraction pathways were found in both brain regions. Although muscle contraction pathways are not expected in brain tissue samples, the overlapping differentially expressed genes were mostly ion channels and signalling cascade proteins also highly relevant for neurons. Figure 5.3 is an example pathway visualization for a pathway that has a high Z-score in both tissue types, Microglia Pathogen Phagocytosis Pathway (*29*).

Pathway-based network construction

From the 904 pathway models in the WikiPathways and Reactome collection, 860 pathways contained 27,410 unique interactions. On average, a pathway contained 35 interactions (min = 1, max = 510, median = 22). Interestingly, 3,264 interactions occur multiple times but only 2,103 interactions are present in more than one pathway. As an example, one of the highest occurring interactions is the complex binding of the three subunits of the IκB kinase complex which plays an important role in the propagation of cellular response to inflammation (*30*) and is present in 25 different pathways.

The resulting network consists of 48,639 nodes and 106,137 edges. The network consists of one major component (46,756 nodes) and 427 smaller components with each less than twenty nodes. The network contains 8,643 gene products, 2,704 metabolites and 9,882 complex / group nodes. Most common interaction types are directed interaction (13,572), complex / group participation (5,298), catalysis (4,787), inhibition (1,185) and conversions (896).

Active module analysis

Active modules were calculated using the jActiveModules app. The top five modules with the highest active paths scores were identified for both comparisons, frontal and temporal cortex. The modules for frontal cortex contained between 300-350 nodes and 560-1,020 edges. The top modules for temporal cortex tended to be smaller ranging from 230-290 nodes and 450-1,000 edges. Figures 5.4 and 5.5 show the highest-ranked module for frontal and temporal cortex, respectively. Gene expression changes are visualized as node color and significance is indicated by the node border color. All modules only contained gene products and no metabolites were found. The complete submodule analysis results for both datasets can be found in Supplementary Data 1 (zip file containing two Cytoscape session files).

The highest ranked active module for frontal cortex contains 303 nodes (79 interactions and 224 gene products) and 568 edges, see Figure 5.4. The subnetwork contains eight significantly down-regulated genes (blue rounded rectangles) including two F-Box genes, *FBOX32* and *FBXO9*, involved in phosphorylation-dependent ubiquitination. The subnetwork contains five significantly up-regulated genes (red rounded rectangles) with diverse roles. The identified hubs in the active module network of frontal cortex are two gene products not measured in the dataset, *RPS27A* and *UBA52*. Both are involved in protein degradation via 26S proteasome, ubiquitination, translation and DNA excision repair. In the central part of the network, the ribosomal proteins including *RPL14*, *RPL29* and *RPL3* form a cluster. This cluster is connected via *PPP2CA* and *PPP2R1A*, two phosphatases involved in cell cycle, DNA replication and transcription, to a cluster of centrosomal proteins including *CEP78*, *CEP57* and *CEP131*. The module combines interactions from 47 unique pathways (Supplementary Table 4) including class I MHC mediated antigen processing and presentation (WP3577), nonsense-mediated decay (WP2710), cell-cycle related pathways (WP1859, WP1775, WP1858, WP2772), and eukaryotic translation elongation and initiation (WP1811, WP1812).

The highest ranked active module for temporal cortex contains 238 nodes (84 interactions and 154 gene products) and 457 edges, see Figure 5.5. The module partially overlaps with the module found for frontal cortex. The module contains 24 significantly down-regulated genes (blue rounded rectangles) including several ubiquitin conjugating enzymes (*UBE2E1*, *UBE2E3*) and translation initiation factors (*EIF4A2*, *EIF4H*, *EIF4G2*). Only five significantly up-regulated genes are found in the subnetwork (red rounded rectangles) but the distance between them is large. This subnetwork contains similar hub nodes as in the frontal cortex subnetwork including *RPS27A*, *UBA52* and *PPP2R1A*. Additionally, *NCBP2* and *NCBP1*, proteins involved in RNA processing, play an important role in the subnetwork. The module combines interactions from 51 unique pathways (Supplementary Table 5) including transcription / translation (WP1889, WP1906, WP1812), cell cycle (WP1859, WP1775, WP4109), and immune response (WP3577, WP2658) related processes.

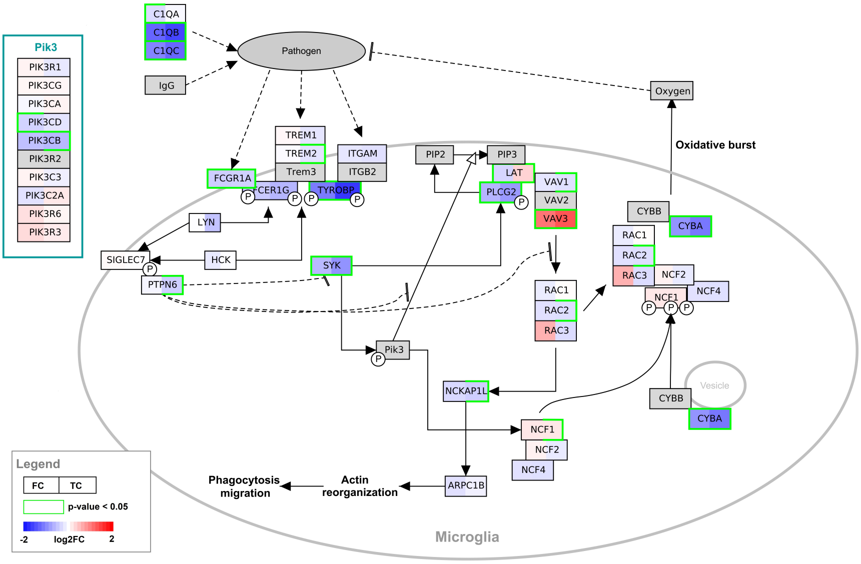


Figure 5.3: **Visualization of the frontal and temporal cortex gene expression on the Microglia Pathway Phagocytosis Pathway.** In the left half of the gene boxes, the gene expression change in the frontal cortex is shown. In the right half of the gene boxes, the gene expression in the temporal cortex is shown. The blue colors represent down-regulation of the gene in Rett syndrome patients (negative log2 fold change), while the red shades are for the up-regulated genes. The darker the color, the stronger the effect. Green borders indicate significance of the change (p-value < 0.05). Grey colored nodes are not annotated or measured in the dataset.

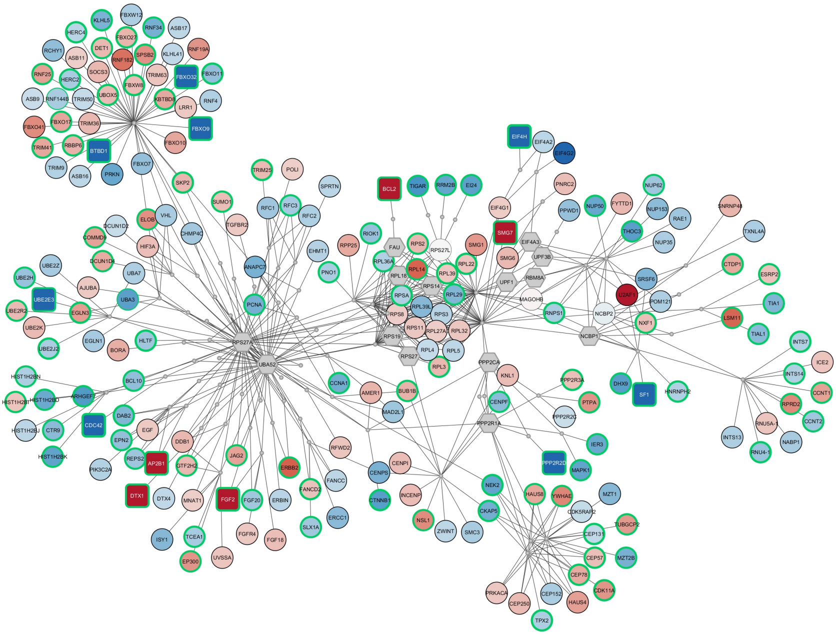


Figure 5.4: **Top-ranked active module for frontal cortex data.** The subnetwork contains 303 nodes and 568 edges. It contains 13 significantly changed genes (rounded rectangles) when applying the same cutoff as for enrichment (absolute fold change > 1.5). Other measured gene products are circular nodes. Blue fill color indicates down-regulation while red indicates up-regulation. The darker the color, the stronger the effect. Gray hexagons are gene products not measured in the data set. The very small, gray nodes represent interaction nodes. These were combined from 47 different pathways with none of the pathways providing more than six interactions.

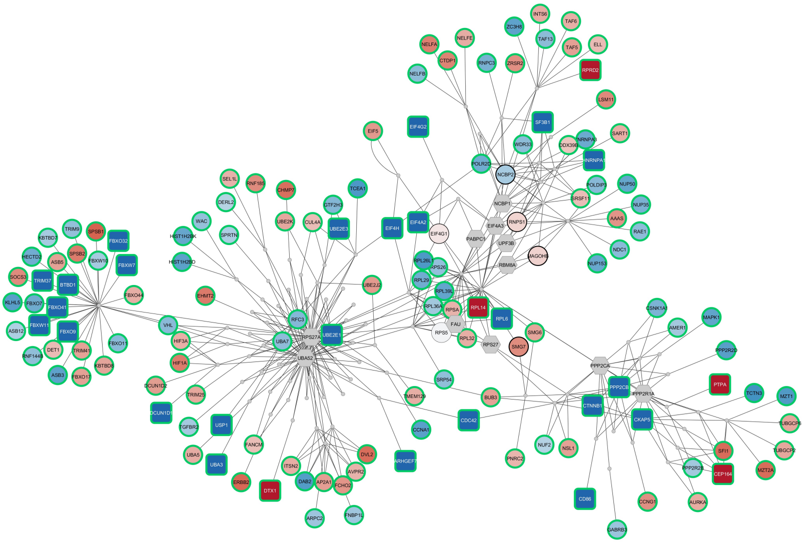


Figure 5.5: **Top-ranked active module for temporal cortex data.** The subnetwork contains 238 nodes and 457 edges. It contains 29 significantly changed genes (rounded rectangles) when applying the same cutoff as for enrichment (absolute fold change > 1.5). Other measured gene products are circular nodes. Blue fill color indicates down-regulation while red indicates up-regulation. The darker the color, the stronger the effect. Gray hexagons are gene products not measured in the data set. The very small, gray nodes represent interaction nodes. These were combined from 51 different pathways with none of the pathways providing more than six interactions.

5.4 Discussion

MECP2 is a multifunctional protein which is involved in several transcriptional inhibitory and activational processes. MECP2 was generally regarded as a repressor, however its role as genetic activator has also been confirmed (*31*). In previous studies, a loss of function in MECP2 due to a mutation has been found to influence a variety of pathways and biological processes, including pathways related to not only neuron development and function, but also to the immune system, transcription and translation related processes (which were identified mainly by transcriptome analysis, (*13–15*, *32*)). The affected pathways identified with our study closely match the results previously found by (*15*), in which human brain tissue data of Rett syndrome patients (published by (*33*)) was analysed. The expression of the MECP2 protein itself is not significantly affected in this dataset (minor, insignificant down regulation (log2 fold change of -0.1) in both brain regions).

The original study by (*16*) from which the dataset analysed in this paper was acquired, focused on the significant down-regulation of certain complement system factors in Rett syndrome (C1QA, C1QB, C1QC). Complement system factors are produced generally in liver, however their expression was also found to be changed in stimulated microglia. Furthermore, there is emerging evidence that C1Q factors are involved in several non-immunogenic activities, such as synaptic pruning in microglia (*34*).

As expected, our pathway and GO analysis revealed a substantial number of immune system related pathways to be affected in Rett syndrome frontal and temporal cortex tissue samples. Inflammatory processes have been identified previously in Rett syndrome patients, mouse models and in vitro systems, and are suspected to contribute to the development of Rett syndrome (*15*, *35*). Figure 5.2 shows many of affected pathways in both frontal and temporal cortex, with similar results found by GO analysis. Interestingly, no complement system or transcription / translation related pathways show up (except Microglia Pathogen Phagocytosis Pathway, which includes C1Q factors). Only seven of the 31 pathways found through pathway analysis contribute interactions to the active modules identified for frontal and temporal cortex. The modules mainly contained interactions from transcription / translation and cell cycle related pathways, which were not found with the classical enrichment analysis. These processes were also previously found in transcriptome pathway analysis of Rett syndrome (*14*, *15*). Overall, the regulatory effects of MECP2, especially on DNA maintenance, cell cycle, transcription and translation, is more prominently shown in the active modules, while immune system related responses are more present in pathway analysis. Importantly, the active module approach does not replace analyses like classical enrichment analysis but augments it.

This was the first time the entirety of the WikiPathways knowledgebase, including Reactome pathways, has been used to create a comprehensive human pathway-based network for network analysis of transcriptomics data. Identifying active modules from a large network has some major benefits, such as the easy applicability to any gene expression dataset, ignoring predefined boundaries used in traditional pathway diagrams, and incorporating the relations and overlap between the pathways. Additionally, this method does not require researchers to predefine a certain cutoff since genes are ranked based on their significance.

Some considerations arose when constructing and analyzing the network. For instance, some common metabolites like ATP, ADP or NADH, while biologically necessary, were excluded from the network, since their involvement in a multitude of interactions created links between almost every node. Additionally, this approach is strongly depending on the a priori input of pathway data in terms of coverage and quality. Currently, human pathway databases contain a little over 50% of the protein coding genes (*4*), which is also a probable number for the coverage of metabolites and interactions. Pathway models generally contain information about directionality of the interactions. However, available active subnetwork analysis methods only take topology but not directionality into account. This could strongly affect the identification of active submodules and would be an important extension of existing algorithms.

The active module discovery approach should be considered as an additional step after classical enrichment analysis. In this study, we used human brain transcriptomics data from a study with Rett syndrome patients, however our approach is not unique to this application or rare diseases. These diseases are by definition less common and often less extensively studied, which may result in lower availability of specific pathway models. Nonetheless, the active module approach succeeds and shows great power for additional discoveries. While rare genetic diseases have the advantage that the causative gene is (usually) known, the resulting downstream consequences can be diverse and affect a variety of pathways. By using pathway models in an integrative network approach, further use of the invaluable resources present in pathway databases is enabled and subnetworks of interest can be retrieved based on the entire body of pathways available. Using Cytoscape allows using all available apps such as the jActiveModules app to analyse our network. Importantly, the complete interaction network of WikiPathways with 48,639 nodes and 106,137 edges can be opened and analysed in Cytoscape, despite of the network to be too large to be visualized. The use of graph databases like Neo4j, which already have connections available to Cytoscape (cyNeo4j app, (*36*)), could be a useful addition to the approach.

Conclusion

Pathway models have proven themselves as powerful tools for biologists to describe and analyse biological processes. The collaboration between the widely-adopted pathway databases WikiPathways and Reactome and the availability of their data in RDF format allowed us to integrate a large number of pathways from both databases into one large network. This enables us to perform advanced network analyses like active submodule identification. By comparing classical enrichment methods with the active submodule identification on a Rett syndrome dataset in two different brain regions, we found that both approaches provided valuable insights into the disease. Importantly, they were strongly complementary and did not show the same results.

Data Availability Statement

The dataset analyzed for this study can be found in the Gene Expression Omnibus: (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE75303).

References

1. R. A. Miller *et al.* Beyond Pathway Analysis: Identification of Active Subnetworks in Rett Syndrome. *Frontiers in Genetics*. 2019. **10**: 59.
2. P. Khatri, M. Sirota, A. J. Butte. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS computational biology*. 2012. **8** (2): e1002375.
3. M. Kutmon, S. Lotia, C. T. Evelo, A. R. Pico. WikiPathways App for Cytoscape: making biological pathways amenable to network analysis and visualization. *F1000Research*. 2014. **3**: .
4. D. N. Slenter *et al.* WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic acids research*. 2017. **46** (D1): D661–D667.
5. A. Fabregat *et al.* The Reactome pathway knowledgebase. *Nucleic acids research*. 2017. **46** (D1): D649–D655.

6. A. Waagmeester *et al.* Using the semantic web for rapid integration of WikiPathways with other biological online data resources. *PLoS computational biology*. 2016. **12** (6): e1004989.

7. A. Rett. On a unusual brain atrophy syndrome in hyperammonemia in childhood. *Wiener medizinische Wochenschrift (1946)*. 1966. **116** (37): 723.

8. R. E. Amir *et al.* Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nature genetics*. 1999. **23** (2): 185.

9. Y. Chunshu, K. Endoh, M. Soutome, R. Kawamura, T. Kubota. A patient with classic Rett syndrome with a novel mutation in MECP2 exon 1. *Clinical genetics*. 2006. **70** (6): 530–531.

10. F. Ehrhart *et al.* Rett syndrome–biological pathways leading from MECP2 to disorder phenotypes. *Orphanet Journal of Rare Diseases*. 2016. **11** (1): 158.

11. J. L. Neul *et al.* Rett syndrome: revised diagnostic criteria and nomenclature. *Annals of neurology*. 2010. **68** (6): 944–950.

12. B. Hagberg, F. Hanefeld, A. Percy, O. Skjeldal. An update on clinically applicable diagnostic criteria in Rett syndrome. Comments to Rett Syndrome Clinical Criteria Consensus Panel Satellite to European Paediatric Neurology Society Meeting, Baden Baden, Germany, 11 September 2001. *European journal of paediatric neurology: EJPN: official journal of the European Paediatric Neurology Society*. 2002. **6** (5): 293.

13. S. Shovlin, D. Tropea. Transcriptome level analysis in Rett syndrome using human samples from different tissues. *Orphanet journal of rare diseases*. 2018. **13** (1): 113.

14. F. Bedogni *et al.* Rett syndrome and the urge of novel approaches to study MeCP2 functions and mechanisms of action. *Neuroscience & Biobehavioral Reviews*. 2014. **46**: 187–201.

15. F. Ehrhart *et al.* Integrated analysis of human transcriptome data for Rett syndrome finds a network of involved genes. *bioRxiv*. 2018. : 274258.

16. P. Lin *et al.* Transcriptome analysis of human brain tissue identifies reduced expression of complement complex C1Q Genes in Rett syndrome. *BMC genomics*. 2016. **17** (1): 427.

17. L. M. Eijssen *et al.* User-friendly solutions for microarray quality control and pre-processing on ArrayAnalysis.org. *Nucleic acids research*. 2013. **41** (W1): W71–W76.

18. M. E. Ritchie *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research*. 2015. **43** (7): e47–e47.

19. C. Leemans, E. Willighagen, A. Bohler, L. Eijssen. BridgeDbR: Code for using BridgeDb identifier mapping framework from within R. *R package*. 2018. **1.16.0**: .

20. M. P. van Iersel *et al.* The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. *BMC Bioinformatics*. 2010. **11** (1): 5.

21. M. Kutmon *et al.* PathVisio 3: an extendable pathway analysis toolbox. *PLoS computational biology*. 2015. **11** (2): e1004085.

22. A. C. Zambon *et al.* GO-Elite: a flexible solution for pathway and ontology over-representation. *Bioinformatics*. 2012. **28** (16): 2209–2210.

23. D. R. Zerbino *et al.* Ensembl 2018. *Nucleic Acids Research*. 2017. **46** (D1): D754–D761.

24. D. Mietchen *et al.* Enabling open science: Wikidata for research (Wiki4R). *Research Ideas and Outcomes*. 2015. **1**: e7573.

25. J. Hastings *et al.* ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic acids research*. 2015. **44** (D1): D1214–D1219.

26. D. S. Wishart *et al.* HMDB 4.0: the human metabolome database for 2018. *Nucleic acids research*. 2017. **46** (D1): D608–D617.

27. P. Shannon *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*. 2003. **13** (11): 2498–2504.

28. T. Ideker, O. Ozier, B. Schwikowski, A. F. Siegel. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*. 2002. **18** (suppl_1): S233–S240.

29. K. Hanspers, D. Slenter, *Microglia Pathogen Phagocytosis Pathway (Homo sapiens)*, https://www.wikipathways.org/instance/WP3937, 2017.

30. H. Häcker, M. Karin. Regulation and function of IKK and IKK-related kinases. *Sci. Stke*. 2006. **2006** (357): re13–re13.

31. M. Chahrour *et al.* MeCP2, a key contributor to neurological disease, activates and represses transcription. *Science*. 2008. **320** (5880): 1224–1229.

32. C. Colantuoni *et al.* Gene expression profiling in postmortem Rett Syndrome brain: differential gene expression and patient classification. *Neurobiology of disease*. 2001. **8** (5): 847–865.

33. V. Deng *et al.* FXYD1 is an MeCP2 target gene overexpressed in the brains of Rett syndrome patients and Mecp2-null mice. *Human molecular genetics*. 2007. **16** (6): 640–650.

34. L. Kouser *et al.* Emerging and novel functions of complement protein C1q. *Frontiers in immunology*. 2015. **6**: 317.

35. C. De Felice *et al.* Rett syndrome: an autoimmune disease? *Autoimmunity reviews*. 2016. **15** (4): 411–416.

36. G. Summer *et al.* cyNeo4j: connecting Neo4j and Cytoscape. *Bioinformatics*. 2015. **31** (23): 3868–3869.

6

An approach to the proposal of drug combination for cancer therapy using a pathway data connectivity approach

Adapted from: R. Miller *et al.*, *An approach to the proposal of drug combination for cancer therapy using a pathway data connectivity approach*, ChemRxiv, Mar. 2022, (https://doi.org/10.26434/chemrxiv-2022-0n8fp).

Abstract

Within the next twenty years, the number of cancer patients is expected to rise by 70%. Current cancer treatments still face several limitations, such as severe side effects and a high incidence of disease recurrence. Drug combination therapies are a promising strategy to achieve higher therapeutic effects while reducing side effects. This new direction in cancer drug research has led to data-driven medicine. To predict whether certain drugs would have a synergistic effect when combined, the DREAM Challenge coordinators released data for thousands of experimentally tested drug combinations. The DREAM Challenge served as inspiration for selecting drug combinations that have the potential to be synergistic. We here describe an approach using biological pathway knowledge and applying this to the selected combinations with a previously described mathematical model, the Loewe-Additivity approach. The calculated interaction index (II) served to distinguish between synergistic ($II < 1$), additive ($II = 1$) and antagonistic ($II > 1$). Pre-selection of putative drug combinations was performed prior to synergy prediction based on four case scenarios: 1) two drugs share the same target protein, 2) two drugs share the same pathway, 3) drugs are separated by one degree from two targets or 4) drugs are separated by more than one degree from two targets but act upon the same biological pathway. Results - The first method tried was using a drug synergy prediction method called the Loewe-Additivity model, in which two drugs share the same target and form the initial findings for this paper. The Loewe model acts as a baseline estimation to see if more combinations can be identified using the other methods tested. The remaining methods used were able to find additional drug combinations that were not proposed by the standard Loewe model. Although the additional methods did find additional combinations that would be predicted to be synergistic, a prediction is not a guarantee of success, so validation of the new or novel combinations would be needed to verify their effectiveness. This could be done by comparing our results to known data or against biological assays.

6.1 Introduction

Over the last two decades, the mortality rates associated with cancer have steadily decreased, constituting an overall fall of 23% in 21 years (*2*). This decrease in cancer cases can be explained by better prevention and an earlier diagnosis, together with better and more effective treatment options. Although mortality rates are declining, cancer is still among the leading causes of death and morbidity worldwide, with the number of cases estimated to rise by 70% in the next twenty years, according to the GLOBOCAN project (*3*, *4*).

This shows a pressing need for not only improved preventive and detection methods, but also for treatment options. Cancer treatments such as chemotherapy, immunotherapy and radiation therapy still face several challenges (*5*). Although various single drug pathway-inhibitors have been developed, cancerous cells, specifically found in solid tumors, are known to have a high resistance to drug treatment allowing the disease to recur (*6*). Currently, the most effective strategy to overcome drug resistance in oncology treatments consists of administering a combination of drugs rather than a single drug (*7*), resulting in a multi-target therapy that takes advantage of the multifactorial nature of cancer.

Due to the fact that cancer is caused by varying underlying mutations in different genes, several cellular pathways are affected by its manifestation; this set of pathways is referred to as the hallmarks of cancer (*8*). The pathways include proliferative signaling, angiogenesis, evasion of growth suppressors, invasion and metastasis, resisting apoptosis, inflammation, genomic instability, and unlimited replication potential. Inhibition of one of these key pathways may still allow cancerous cells to survive and adapt to the administered drug therapy. The likelihood of drug resistance decreases when multiple pathways that are known to be relevant in cancer are targeted (*9*, *10*). There still needs to be a consideration of the role of the pathways targeted and their relationship between one another.

The first choice as an alternative to the selection of two drugs that share a common target was to use two drug targets present in the same pathway, because under the standard version of mathematical results, we use two drugs that are targeting the same protein. The rationale to extend this to a whole pathway is that pathways are typically used and drawn to describe a particular biological process. Presumably this means that the protein targets found within a pathway are all part of a related biological process. We can then evaluate these combinations as a baseline to see what sort of predictions of synergy we get from its application. From here we can work on approaches that do not include the same pathway and look at targets that are connected at a distance from each other regardless of their pathway association.

Drug combinations can induce an additive effect; in which the therapeutic outcome is equal to the effect of monotherapy; an antagonistic effect which reduces therapeutic effects or they can lead to a synergistic effect in which the therapeutic effects are enhanced (*11*). Synergistic combinations may offer advantages over mono-drug therapies, because the therapeutic effect in synergistic drugs is greater than simply adding the two drugs' effects together (*12*, *13*). In addition, reduced drug resistance may occur due to synergistic drug combinations, because each individual drug needs a lower dose. Furthermore, a drug compound which normally does not induce an effect on its own could potentially have an effect when in combination with another drug, giving rise to a larger range of possible drug treatments. Due to the large number of possible drug combinations, large and complex data make studying every possible combination difficult. Important developments include the use of genomic data in drug discovery, the sharing of clinical-trial data, as well as the increased availability of data from claims and patient registries (*14*). This new paradigm has been referred to as data-driven medicine (*15*), an area which explores big data, that is, datasets that are too large and complex for traditional data processing applications to work with. Thus, large amounts of data encourage the use of advanced computational analytical methods which enable researchers to take into account a wide set of data beyond those generated in standardized clinical trials and analyzed through classical biostatistics. Harnessing the power of big data can potentially improve success rates in the outcomes of patients receiving treatment as well as enabling pharmaceutical companies to enhance the productivity of their research and development (*16*, *17*). With the use of big data, researchers can pursue more computational approaches to drug development such as the one described in this paper that would be predicted to have a synergistic outcome or not.

But more data does not directly give us new answers. As we see more and more data available to researchers, we need new ways to look at and approach this data to look for meaningful connections. Under this mindset we see companies, organizations, and governments releasing some of their data to crowdsource new approaches to using their data. AstraZeneca, for example, established a partnership with the European Bioinformatics Institute, the Sanger Institute, Sage Bionetworks and the DREAM community and released approximately 11.5k experimentally tested drug combinations measuring cell viability in several cancer cell lines and challenged the public domain to explore potential synergistic drug combinations (*18*). The question remains, however, how this data can be used to get new insights in the applicability of drug combinations. Combinations of drugs are used in medicine to treat ailments that are complex diseases, such as cancer, and can have synergistic effects (*19*). Drug synergy is defined as the effect of a combination of drugs that is greater than summing the effects of the individual drugs together. There are different models for predicting effects based on component-based approaches, with the independent model (IA) and the concentration addition model (CA) being described by Tanaka and Tada (*20*). The CA model is for similarly acting chemicals, while the IA model is used to describe chemicals with different modes of action. The CA model is generally considered to be a better predictor of the synergy than the IA model in the case of a mixture of many substances (*20*). The Loewe-Additivity model is an example of a CA model and follows the model of similarly acting substances if they target the same targets and processes as defined by the equation

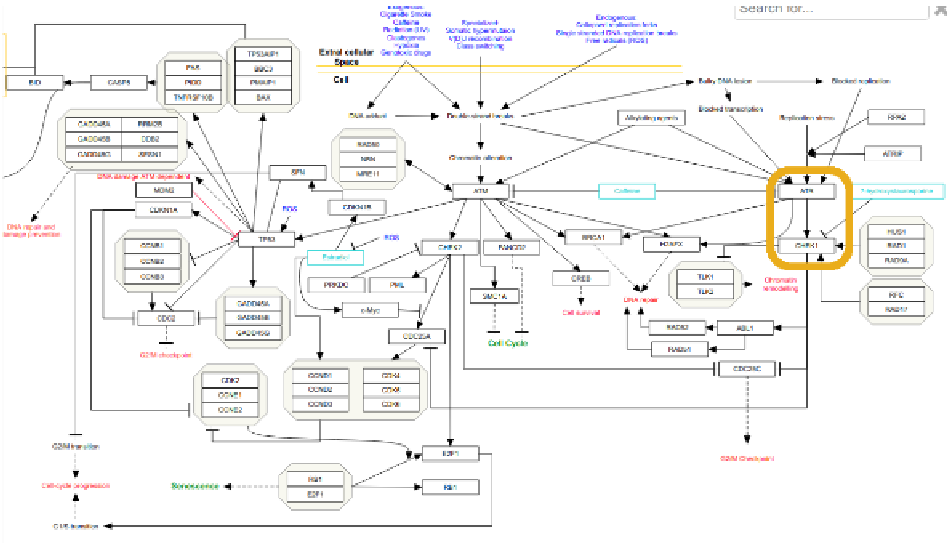


Figure 6.1: Example interaction from DNA damage response pathway (wikipathways:WP707) from WikiPathways. The line and the arrow between the gene product nodes is the interaction and directional information captured in the WikiPathways RDF.

$Synergy_{Loewe} = \frac{d1}{D1} + \frac{d2}{D2} = 1$ for additivity, where if $Synergy_{Loewe} < 1$ is synergistic and $Synergy_{Loewe} > 1$ is antagonistic with d1 and d2 represent doses that have an effect when in combination and D1 and D2 have an effect when used alone.

Therefore, starting from the ideas of additivity of drug synergy, and the knowledge how genes are related via their biological processes, the aim of this research paper is to propose and select drug combinations that are connected biologically, either by process type or by common branch connectivity. This could facilitate the identification of potential drug combinations to investigate for synergy without already having measured information about the drugs' potential synergy previously. The approach being examined is that drug combinations can be identified with network connectivity and extend an already established mathematical model, the Loewe-Additivity mathematical model. The principle of Loewe-Additivity model is that two drugs that work on the same drug target can be calculated whether the combination is additive, antagonistic, or synergistic (*21*). The standard Loewe-Additivity model has limitations in that two drugs must share a common target (see Figure 6.3A).

Drug-target information that was already supplied by the DREAM challenge coordinators was the basis for expanding with pathway connectivity data from sources that have information based upon prior knowledge about target interactions with each other. This collection of diagrams are the pathways found on WikiPathways (*22*). Within this collection is information about the connectedness of nodes to form a diagram of biologically related information as seen in Figure 6.1. The nodes in the diagram are representations of gene products, metabolites, proteins, and RNA. This connected diagram supplies information about the interactions involved. With the inclusion of interaction information from WikiPathways, information about the direction in which an interaction is directed is also captured. In this case, this is directional information between "gene product" or "protein" nodes. This allows the data nodes that represent gene products or proteins to be found in a one way direction from the node of interest. The pathway data's semantic representation is in the form of the WikiPathways Resource Description Framework (*23*). For example, in WikiPathways we have interaction directions that indicate that PI3K activates AKT which in turn inhibits the TSC1/TSC2 dimer in the PI3K pathway (Figure 6.2, wikipathways:WP4141), which is an attractive pathway to target in cancer research (*24*). What this directional information permits is finding connections to new gene products or proteins at increasing distances. The pathway data used to find target combinations of increasing distance from each other was the WikiPathways RDF (*23*). The WikiPathways RDF was used because it contains the semantic information for the pathway diagrams at WikiPathways.org. The RDF being the semantic representation of the nodes, interactions, and metadata associated with the pathways. The WikiPathways RDF specifically

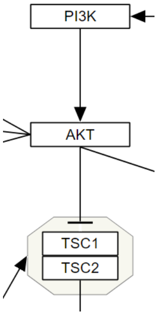


Figure 6.2: Example directional interaction of PI3K activating AKT which in turn inhibits the TSC1/TSC2 dimer found in WikiPathways (wikipathways:WP4141)

captures information about interactions and which elements come before or after another (*25*). Because the RDF has this connection information, the WikiPathways RDF makes a logical choice when one wants to traverse a pathway from one point to another. In this case, it would be to look for drugs that share the same pathway or for paths within a pathway that share drug targets.

The WikiPathways SPARQL endpoint (sparql.wikipathways.org) is a public resource where users can construct queries to retrieve annotated information about the targets (*22*). In this case, the information retrieved was first to see if a certain distance between targets, in relation to each other, maintained enough biological relations. This was extended to include drug combinations that shared a common pathway and should yield more results. Finally, in an attempt to observe what happens when the network extends beyond the immediate biological processes, a network was constructed to connect drugs across pathways via directional queries.

Table 6.1: Example of mono-drug therapy measures of two drugs and their targets.

	0	0.75	2.5	7.5	25	75	(=Agent2)
0	100	97.8	99.7	94.7	95	92.9	
0.01	96.5	NA	NA	NA	NA	NA	
0.03	97.6	NA	NA	NA	NA	NA	
0.1	95.6	NA	NA	NA	NA	NA	
0.3	88.8	NA	NA	NA	NA	NA	
1	79.4	NA	NA	NA	NA	NA	
(=Agent1)		NA	NA	NA	NA	NA	
Agent1	AKT	NA	NA	NA	NA	NA	
Agent2	ADAM17	NA	NA	NA	NA	NA	

6.2 Methods

6.2.1 Drug-Target Information

There were several methods and approaches that were used to propose potentially synergistic drug combinations. As part of the DREAM challenge (*26*), the coordinators provided information about which drugs were used on 85 different cell lines and the targets of these drugs. It was the drug target list that was used for the potential combinations in the different approaches. The list of potential targets is found in the "targetlist.csv" file at https://github.com/RyAMiller/DrugTargetSynergy. Table 6.1 shows AKT 1 concentrations (μm) identified in the first column. ADAM17 (μm) is identified in the first row. The numbers in the second row and column are cell survival rates at the corresponding concentrations. The targets used were the Agent1 and Agent2 values from these tables.

6.2.2 Mathematical model for determining a drug combination's synergy

The Loewe-Additivity model is the mathematical starting point used to predict drug synergies (27). As part of the challenge data, we have information about the drug compounds and their targets. Therefore we can calculate whether the combination is synergistic or not (27). Since the conditions for this model are already well accepted and can be used to calculate drug synergies, this does not give many new or interesting combinations that have yet to be explored for potential synergistic effects. In this case, the established model is considered the standard that will be compared against the new models proposed. The decision was then made to try and extend the model to gene products and proteins that are biologically related, in order to find new combinations that could potentially be synergistic, without just grabbing every permutation of two drugs from the drugs used. This extension is done using biological databases such as WikiPathways, technologies like the Resource Description Framework (RDF), and SPARQL Protocol and RDF Query Language (SPARQL).

Thus the first step was to use a SPARQL query to the WikiPathways endpoint with the query being available in the GitHub repository. This returned a table of source and target proteins that are associated with the compounds used for synergy analysis. The tables were then imported into a spreadsheet program and the concatenate function was used to return pairs of drug combinations to be evaluated for drug synergy separated by a comma.

The first query proposes the drug combination synergies under the assumption of the Loewe-Additivity Model, that two potential drugs both targets act on the same protein. In order to expand this assumption and assuming that proteins in the same pathway are involved in similar biological function, we created a list of synergistic drugs that was a combination of the standard two drugs, one target approach and added a list of two drugs that share one pathway.

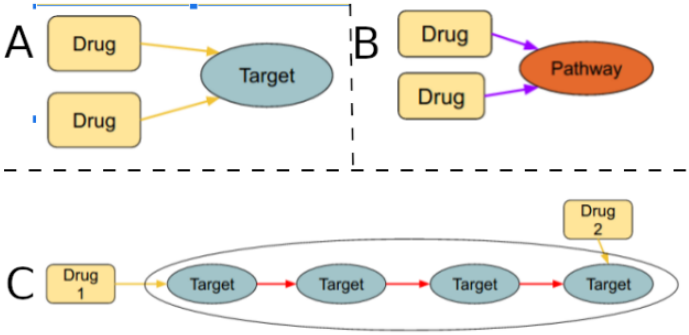


Figure 6.3: The first case for combination selection is done by using the standard conditions for the Loewe-Additivity model. It predicts synergy of the combination based on the two drugs sharing the same target. B: To expand the selection criteria, if both members of a drug combination are present in a common pathway then the combination is selected and sent for synergy calculations. C: The connectivity of the targets to each other is represented in this selection method. If there is a directed interaction between the targets and the drugs share a common interaction stream, then they are selected and are sent to be evaluated for synergy.

Then directional queries were used to say that the two drug targets must be directly upstream or downstream of each other and the results can be seen from the Java script found on the GitHub project page. A similar list of drug target combinations was generated by concatenating drug and target columns and this list was added to the standard two drugs, one target list of drugs to return drug synergy predictions based upon two targets being involved in the same path and can be traced directly from one target to another.

6.2.3 Application Using Loewe Model: Synergy using standard Loewe-additivity model

The first tests were based on using a standard Loewe approach to propose synergy (i.e. two drugs, same target) (see Figure 6.3A). Each of the selected combinations from the scenario where two drugs share the same target were synergistic based upon the mathematical model of the Loewe-Additivity model. These combinations were our baseline list of potentially synergistic drug combinations. Other approaches to propose combinations all include this original list of combinations.

6.2.4 Alternative 1: Drug combination shares a common pathway

One crucial assumption that is made is that a particular family of targets that is identified, includes all potential members of that family in the number of possible combinations. Since the pathway diagram describes a particular process or processes, the RDF can also be used to see if two targets are part of the same pathway and if they are part of the same process then the combination could be potentially synergistic. The simplest list of combinations to construct was the list that has the two drugs sharing a common pathway (Figure 6.3B). This has to be done using the drugs' targets rather than the drugs themselves, because gene products and proteins are the entities that are most represented in WikiPathways' pathway diagrams. The way the annotations are represented in the WikiPathways RDF allows the use of a SPARQL query to determine if two drugs share a common pathway. The query takes a list of the targets and attempts to match the targets to pathways. A query result is returned when all the conditions of the query are met, mainly that the query finds any two targets that are present in any pathway and filters targets to match those identified targets. This again allows the creation of target combination lists. The targets are then associated with their specific drug. This list of combinations plus the list from the Loewe model our the next list of potentially synergistic drug combinations.

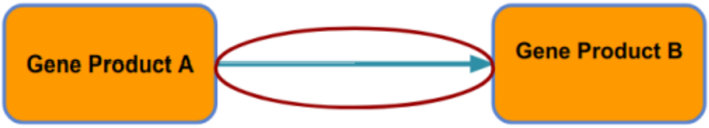


Figure 6.4: Dark red circle highlights a basic interaction. The arrow points from one direction to another and implies directional information about the interaction.

6.2.5 Alternative 2: Combinations share a common path within a pathway

The important aspect of the WikiPathways RDF, for this application, is that it contains information about connectivity of gene products and proteins. This can be applied using SPARQL queries designed to retrieve data about adjacent nodes. The queries used were implemented using Java since the size of the queries required will cause the WikiPathways SPARQL endpoint to time out. The Java code and queries can be found on GitHub. The RDF also contains information about the direction of the interaction (Figure 6.4). Having information about the start and end point of an interaction allows the user to have knowledge of direction of the interaction. In other words, in which direction the arrow is pointed. First the queries started with one of the targets from the challenge data and returns all the targets that are n interactions downstream of the original target list, this can be seen in Figure 6.3C. In this case, n steps are 4 downstream checks. Since the starting point must be one from the list of targets, we would find all combinations of two targets that are associated with the drugs in both the up and down direction. If starting at the origin point of targets is recognized in combination with another target gene within the 4 steps, it is used in the model to test if it is a potentially synergistic combination of targets that can be used against the drugs. This should generate a list of targets that is closely related biologically but adds some freedom to the original Loewe model.

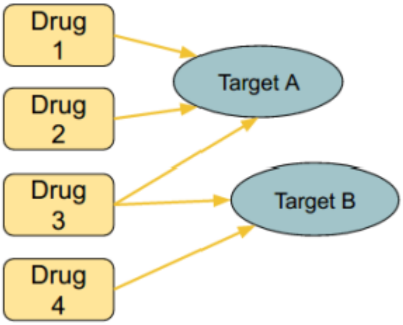


Figure 6.5: Selection criteria is based upon a more indirect connection between targets that have drugs with multiple effective targets. In this case Drug 1 and Drug 4 would be selected as a potential combination to calculate drug synergy.

6.2.6 Alternative 3: Network extended beyond pathway boundaries

The final step to test the limits of the mathematical model, was to create networks of pathways with the target proteins and compare the list of drug combinations to the lists created in the previous approaches, as seen in Figure 6.5. This was done by creating a network in Cytoscape (*28*) of drug and pathways. The Python library, NetworkX, was then used to calculate the shortest path for all different possibilities across the network that had two drugs from the challenge data. This list of edge connections is what NetworkX requires in order to calculate distance across a particular network. Since the drug targets are the elements linking the pathways together, the minimum distance to give a potentially unique solution was five nodes on the network away from each other and not just another case of two targets sharing a common pathway as done before.

6.2.7 Repository

A GitHub repository at https://github.com/RyAMiller/DrugTargetSynergy contains the scripts used for this project. This repository contains java scripts with SPARQL queries for directional interactions and the tables for drug targets. The lists of drug combinations files are also found in this repository as well as the CytoScape networks used in this manuscript.

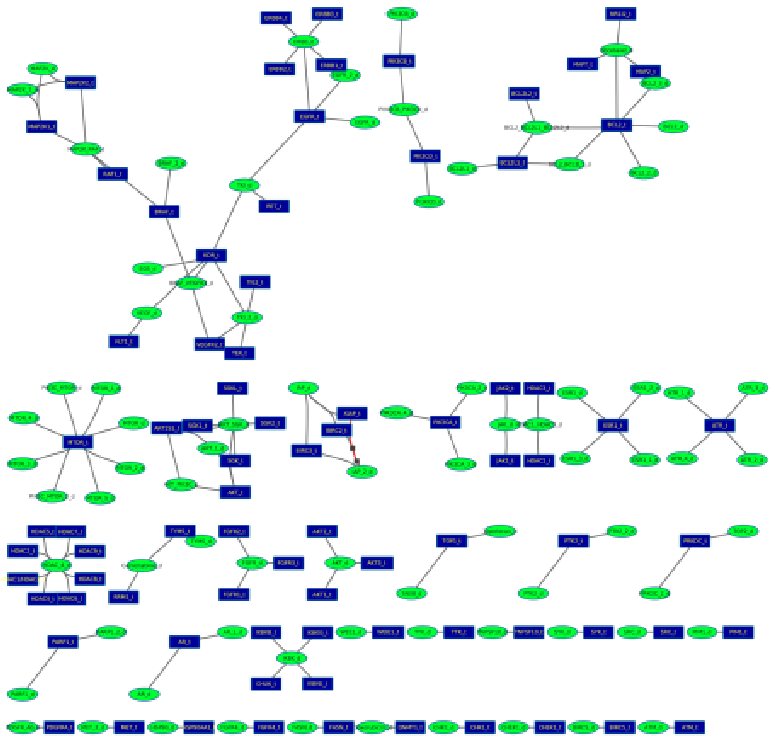


Figure 6.6: Drugs that share a common target network. This is a simple network based upon the drugs and their targets. Drugs are connected by the targets with which they are involved.

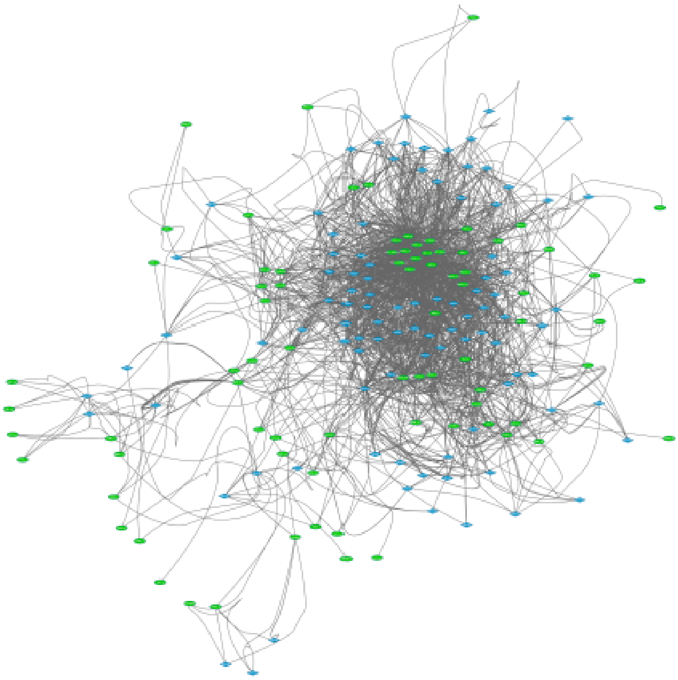


Figure 6.7: Drug/Pathway Network. Expanding the network to also include drugs that are connected via the same pathway.

6.3 Results

6.3.1 Workflow

The standard Loewe model and the other methods were completed as described above. Based upon the models, lists of potentially synergistic combinations were created.

6.3.2 Loewe model network results

The selection of drug combinations that are applied to the model in different ways. In the common model for the Loewe-Additivity model, the graphical representation looks like Figure 6.3A where two drug compounds have a predicted synergy if they interact with a common target. In Figure 6.6, the network is created for the drug-target combinations. This network shows that there are 132 edges and 161 nodes. This shows what kind of connectivity there is with a well studied model like the Loewe-Additivity model. This yields 328 drug combinations that are possibly synergistic.

6.3.3 Targets share the same pathway

The Loewe-Additivity model is already well described so in Figure 6.7 the idea was to use the same mathematical model but apply it to a combination of two drugs that shared a common pathway and then create a network of the drug-pathway combinations. This shows a much larger network that has 177 nodes and 1811 connections. This increased connectedness, compared to the standard approach, also gives an increase in the number of potential combinations to send to the scoring algorithm. This also greatly increased the number of combinations that are possibly synergistic to 1089.

6.3.4 Directional information between targets

Just because two drugs share a common pathway does not mean that the two individual drugs are even in the same arm of the pathway with common effects to be seen downstream. In an attempt to have a set of combinations only with two drugs sharing a common interaction stream, a different approach was specifically done to have only combinations which share common effects if targeted (Figure 6.8). This network has 6449 edges and 648 nodes. At a limit of 4 interactions

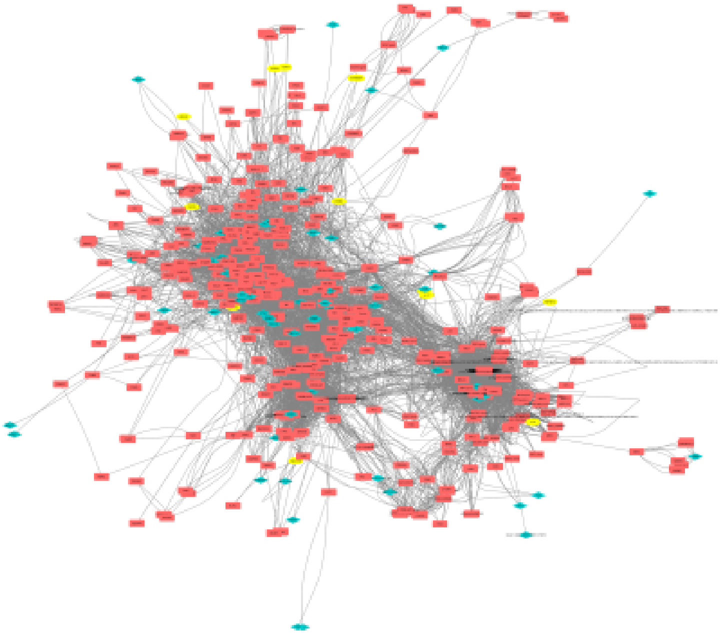


Figure 6.8: Directionality from a starting point (defined as targets) and using directional information from WikiPathways' pathways four interactions removed from the starting point.

allowed between the two drug targets, the number of pottential synergies is 1089. The directional method produced a similar number of potential combinations as the shared pathway approach but in the directional method, there is increased confidence about the targets having a common path that connects the two and more confidence in their relation to each other biologically.

6.3.5 Network extended beyond pathway boundaries

The final attempt at determining possible combinations would be to look for combinations that are only connected by interaction informa-

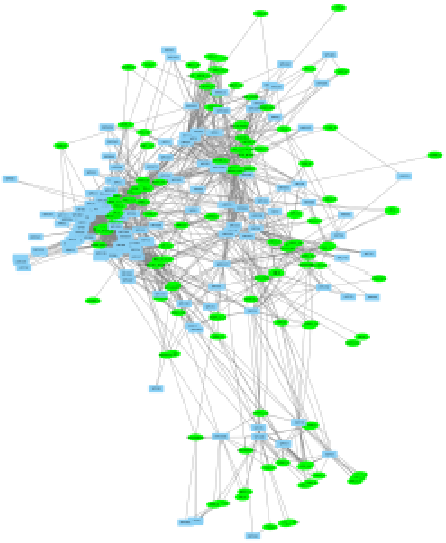


Figure 6.9: Indirect connections. In this case the network was constructed so that the pathway boundary was not a limitation to connections. A gene product or protein can be connected with a pathway but the pathway shares something in common with the next pathway.

tion from WikiPathways but without regard to pathway boundaries (Figure 6.9). The idea of this last approach was to find new potential combinations by extending the network with connections to other pathways that are not associated with the gene products or proteins that are the identified targets. This type of expansion, however, is likely to be much too far removed from the biological explanation using the mathematical model. This did turn up a far greater number of predictions at 7545 synergistic predictions, with 2223 edges and 315 nodes. This means that the network becomes more and more connected the further the network is expanded. If extended too far, then the entire WikiPathways graph will connect all pathways and nodes to each other and almost every combination of drug targets is possible.

6.4 Discussion and Conclusion

This manuscript proposes a method to create combinations of drugs that could potentially be synergistic in nature by looking at process and connectivity information for their targets in order to screen combinations of drugs that are acting upon a common mechanism within a biological system. The specific challenge identified by the coordinators that was undertaken was to not use training data to try and make predictions for drug combinations that are synergistic. This was done using an existing mathematical model, but trying to give the model parameters outside its normal range. The approaches that were used were done in order to use existing biological knowledge about how potential drug targets are connected and use this connectivity information with the assumption that the combinations are working on similar biological processes. Retrieving specific interaction about target connectivity information acquired from WikiPathways allowed us to select combinations for predictions based upon publicly known pathway data about connectivity of gene products that share the same biological course. It is this connectivity that allows this approach to be rooted in accepted biological principles. The project was successful in proposing drug combinations based only on biologically relevant connections from pathway data, and the application of the Loewe mathematical model to conditions outside the model's normal parameters, unique combinations can be identified that are overlooked by the standard Loewe model.

What we saw as the networks grew was that as the networks grew bigger, it allowed for more permutations of drug combinations. If the drug targets were too distant from each other, then almost all permutations of the two drug combinations were possible. One way to help combat this is to use directional information about the connectivity between the two targets to keep combinations in the same branch with the assumption that regulation of one of the targets has a downstream effect upon the other target and targeting the processes at different points has the potential to have a greater than standard effect at the drug dosage. The distance between elements cannot grow too large as the possibilities of alternative paths of expression becomes greater.

While the approach described in this paper was successful in using connectivity information to link the drugs to the processes that they are targeting and finding combinations that affect similar processes we see the limit of the standard Loewe mathematical model becomes a concern when we start to work outside the original specifications of the model when trying to predict drug synergy. That is when we start applying the model to conditions different from two drugs sharing the same target. The Loewe model makes the assumption that two drugs are both acting upon a common target. A more appropriate approach would be to use a Generalized Concentration Addition (GCA) model that handles common processes better than a standard CA approach like the Loewe-Additivity Model. In order to further develop and refine this method of combination prediction for drug synergy, it would be interesting to look at a mathematical model more suited for our combinations. One technique that would be possible would be to use the GCA model since it improves upon the more basic CA model (*20*, *29*). In follow up experiments, we would prefer to use find transcriptomics data to see if the pathways which contain the targets are also affected by the drugs of interest. Labib *et al.* (*29*) performed an interesting study to predict hazards associated with complex mixtures of polycyclic aromatic hydrocarbons. They did this by using the GCA model in combination with a pathway benchmark done. They first selected pathways relevant for cancer formation and then calculated the dose-response for each individual compound and pathways using CA, IA, and GCA models and compared their predictions to observations. A similar approach could be used as a followup for our combination predictions and would require transcriptomics data for the drugs at various concentrations.

The implication of the identified combinations is that the predictions are potentially synergistic combinations. Real world effectiveness of the drug combination predictions has yet to be determined, but present combinations of drugs that target proteins that are biologically connected to each other through pathways and identify different protein targets at multiple points in pathways. If the team were to repeat this experiment again, we would use a mathematical model, such as a GCA model, that follows the parameters for our specific data and approach. The biological information and how genes are connected does impact synergy. This is shown in the application of methods that share a common route through the pathway and the predicted combinations list is generated.

Acknowledgements

This manuscript was made possible and inspired by a project that was part of the work of students from the Maastricht Science Program for the DREAM challenge. The students' work was foundational to this paper and would not have been possible without the assistance of Moritz Schlichting and the rest of the students.

References

1. R. Miller *et al.*, *An approach to the proposal of drug combination for cancer therapy using a pathway data connectivity approach*, ChemRxiv, Mar. 2022, (https://doi.org/10.26434/chemrxiv-2022-0n8fp).
2. R. L. Siegel, K. D. Miller, A. Jemal. Cancer statistics, 2016. *CA: A Cancer Journal for Clinicians*. 2016. **66** (1): 7–30.
3. S. Antoni, I. Soerjomataram, B. Møller, F. Bray, J. Ferlay. An assessment of GLOBOCAN methods for deriving national estimates of cancer incidence. *Bulletin of the World Health Organization*. 2016. **94** (3): 174–184.
4. J. Ferlay *et al.* Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *International Journal of Cancer*. 2014. **136** (5): E359–E386.

5. M. Nikolaou, A. Pavlopoulou, A. G. Georgakilas, E. Kyrodimos. The challenge of drug resistance in cancer treatment: a current overview. *Clinical & Experimental Metastasis*. 2018. **35** (4): 309–318.

6. J. Zugazagoitia *et al.* Current Challenges in Cancer Treatment. *Clinical Therapeutics*. 2016. **38** (7): 1551–1566.

7. I. Bozic *et al.* Evolutionary dynamics of cancer in response to targeted combination therapy. *eLife*. 2013. **2**: .

8. D. Hanahan, R. A. Weinberg. The Hallmarks of Cancer. *Cell*. 2000. **100** (1): 57–70.

9. X. Hu, Y. Xuan. Bypassing cancer drug resistance by activating multiple death pathways – A proposal from the study of circumventing cancer drug resistance by induction of necroptosis. *Cancer Letters*. 2008. **259** (2): 127–137.

10. K. S. Smalley *et al.* Multiple signaling pathways must be targeted to overcome drug resistance in cell lines derived from melanoma metastases. *Molecular Cancer Therapeutics*. 2006. **5** (5): 1136–1144.

11. J. J. Peterson, S. J. Novick. Nonlinear Blending: A Useful General Concept for the Assessment of Combination Drug Synergy. *Journal of Receptors and Signal Transduction*. 2007. **27** (2-3): 125–146.

12. D. S. Wald, M. Law, J. K. Morris, J. P. Bestwick, N. J. Wald. Combination Therapy Versus Monotherapy in Reducing Blood Pressure: Meta-analysis on 11,000 Participants from 42 Trials. *The American Journal of Medicine*. 2009. **122** (3): 290–300.

13. T.-C. Chou. Drug Combination Studies and Their Synergy Quantification Using the Chou-Talalay Method. *Cancer Research*. 2010. **70** (2): 440–446.

14. N. Szlezák, M. Evers, J. Wang, L. Pérez. The Role of Big Data and Advanced Analytics in Drug Discovery, Development, and Commercialization. *Clinical Pharmacology & Therapeutics*. 2014. **95** (5): 492–495.

15. N. H. Shah, J. D. Tenenbaum. The coming age of data-driven medicine: translational bioinformatics' next frontier. *Journal of Informatics in Health and Biomedicine*. 2012. **19** (e1): e2–e4.

16. W. Raghupathi, V. Raghupathi. Big data analytics in healthcare: promise and potential. *Health Information Science and Systems*. 2014. **2** (1): .

17. I. Hernandez, Y. Zhang. Using predictive analytics and big data to optimize pharmaceutical outcomes. *American Journal of Health-System Pharmacy*. 2017. **74** (18): 1494–1500.

18. J. Dry, *AstraZeneca-Sanger Drug Combination Prediction DREAM Challenge*, https://www.synapse.org/#!Synapse:syn4231880/wiki/235645, 2016.

19. K. R. Roell, D. M. Reif, A. A. Motsinger-Reif. An Introduction to Terminology and Methodology of Chemical Synergy—Perspectives from Across Disciplines. *Frontiers in Pharmacology*. 2017. **8**: .

20. Y. Tanaka, M. Tada. Generalized concentration addition approach for predicting mixture toxicity. *Environmental Toxicology and Chemistry*. 2016. **36** (1): 265–275.

21. W. E. Delaney, H. Yang, M. D. Miller, C. S. Gibbs, S. Xiong. Combinations of Adefovir with Nucleoside Analogs Produce Additive Antiviral Effects against Hepatitis B Virus In Vitro. *Antimicrobial Agents and Chemotherapy*. 2004. **48** (10): 3702–3710.

22. M. Martens *et al.* WikiPathways: connecting communities. *Nucleic Acids Research*. 2020. **49** (D1): D613–D621.

23. A. Waagmeester *et al.* Using the Semantic Web for Rapid Integration of WikiPathways with Other Biological Online Data Resources. *PLOS Computational Biology*. 2016. **12** (6): e1004989+.

24. K. D. Courtney, R. B. Corcoran, J. A. Engelman. The PI3K Pathway As Drug Target in Human Cancer. *Journal of Clinical Oncology*. 2010. **28** (6): 1075–1083.

25. R. A. Miller *et al.* Understanding signaling and metabolic paths using semantified and harmonized information about biological interactions. *PLOS ONE*. 2022. **17** (4): e0263057.

26. M. P. Menden *et al.* Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. *Nature Communications*. 2019. **10** (1): 2674.

27. M. Goldoni, C. Johansson. A mathematical approach to study combined effects of toxicants in vitro: Evaluation of the Bliss independence criterion and the Loewe additivity model. *Toxicology in Vitro*. 2007. **21** (5): 759–769.

28. P. Shannon. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*. 2003. **13** (11): 2498–2504.

29. S. Labib *et al.* A framework for the use of single-chemical transcriptomics data in predicting the hazards associated with complex mixtures of polycyclic aromatic hydrocarbons. *Regulatory Toxicology*. 2016. **91** (7): 2599–2616.

7 Discussion

This thesis explores to what extent it is possible to use connectivity and directional information of a resource such as WikiPathways (*1*) to increase understanding in biology. A biological organism is a system of things working together, and the elements of a system rarely act outside the system as the system cooperation is necessary for a healthy organism. This means that changes to individual parts of the system have effects on other parts of the system. These make it advantageous to look at not only individual parts of a system, the genes, proteins, or metabolites, but also studying the entire system for changes (*2*). Studying changes at a system or organism level allows for studying effects and conditions that are more complicated than a single element or relationship such as diseases (*3*), changes caused by drugs in pharmacology (*4*), or holistic network approaches in cancer (*5*). The higher level view of the system demonstrates the importance of this work to make pathway data and connectivity data available to the scientific community and shows applications about how it can be used.

How the parts of a system are connected to one another in systems biology is an important aspect of understanding how an organism works together to succeed in life. The directional nature of the edges connecting pathway elements reflect the biological influence of one part on the other, whether that is inhibiting another part or if it is stimulating it. This thesis has shown that not only knowing the direction but also understanding the type of influence can also be useful to understanding. Whether it is inhibition, stimulation, conversion, catalysis, transcription-translation, or more general directional arrow types of interactions which are supported by resources like WikiPathways, they also give additional information to aid in the understanding of how all the system's parts work together. Knowing if, for example, that an entity inhibits the expression of another gene or protein or if it stimulates its production (6), is important contextual information that further explains an organism or a specific biological function. It is this understanding of these parts of a biological system such as the direction of an interaction or the type of an interaction that is a consistent theme of this thesis. These biological parts make pathway diagrams more useful via machine-readable format and computational queries, it shows how this furthers the study and understanding of the system as a whole.

7.1 Semantic Pathway Representation

In order to represent pathway diagrams, which are graphical representations of a biological process, we converted the graphical format used for pathways to a semantic representation. For WikiPathways this involved turning the GPML, which is an XML-based format and the output file for pathway diagram saved in PathVisio (7) and turning these GPML pathway files into RDF. An RDF representation allows for query languages like SPARQL to query the data using an appropriate SPARQL endpoint. Although GPML is a semantic format as well it is not a widely used standard that allows for the query of multiple sources with one standard used across many biological databases as RDF does. An endpoint was established for the WikiPathways data that is publically available. An advantage of having an endpoint for the WikiPathways RDF is that it is able to be queried using federated and non federated queries. These federated queries allow for the querying of data from WikiPathways and other resources simultaneously. This means that one query is able to access both WikiPathways and another data source at once, for example, a single query can be made to query WikiPathways and Ensembl (*8*) together. Allowing for the easy integration of data from both sources, which may be missing information. Allowing WikiPathways to provide some information about how things are connected and Ensembl to provide information about specific gene entities. This is the ability to represent pathway diagrams as semantic data that can be queried and retrieved later.

The WikiPathways vocabulary is also defined and focuses on biological meaning as demonstrated in Chapter 2. The RDF uses this vocabulary to provide appropriate predicates for the RDF triples. These vocabularies were necessary to describe diagrams drawn in WikiPathways. Other ontologies are also found in pathways, such as dcterms from the Dublin Core. For WikiPathways, identifiers.org is used to provide resolvable Uniform Resource Identifiers (URIs) (*9*). The semantics of WikiPathways along with federated queries allows for WikiPathways data to be queried via various analysis platforms, like for example, in the Open PHACTS discovery platform. In order for the WikiPathways RDF to be more usable, to a wider audience, we still need to describe how elements are connected to one another within the RDF.

7.2 Pathway Interactions

Understanding the semantics of the pathway is the first need to understand how to use pathway connectivity in the WikiPathways RDF to be able to answer questions in biology. Pathway connectivity is an essential dynamic for studying other conditions. Chapter 3 studied how the semantically harmonized pathways can be used to study the interations. It shows that it is possible to use semantic data from WikiPathways to answer questions related to how pathway elements are connected with each other and what the biological meaning is. This was demonstrated with two use cases: e.g. 1. sphingolipid metabolism and 2. connectivity of MECP2. Sphingolipids are important biological structural molecules that need to be managed by organisms and cells and are used in regulation (*10*). Understanding how metabolites are converted from one form to another by an enzyme catalyzing their reaction is important to understanding the regulatory nature of sphingolipids. We found that such enzymatic reactions can be readily recognized with a query.

The other pathway element studied was the protein encoded by MECP2. MECP2 is a gene that is responsible for suppressing expression of other genes found in nerve cells. When the protein is not functioning properly it can cause problems in the development of mammalian organisms (*11, 12*). MECP2 is linked to rare developmental diseases such as Rett Syndrome (*13*). Understanding how MECP2 influences and is influenced by other elements can give an idea of how the changes in MECP2 can directly influence other parts of the organism's normal biological system. Here, I showed that it is possible to easily retrieve elements upstream and downstream of MECP2 to clearly illustrate how MECP2 influences or is influenced by other elements in the system, but more than one interaction away. This demonstrates how many elements are upstream of MECP2 and how many are downstream of it. A user would want to use this because they can apply this method to any gene or protein of interest and find influences without manually counting individual interactions emanating from or converging on said gene or protein. This captures the semantics of the biological meaning for the pathway.

This chapter also studied how Shape Expressions (ShEx) (*14*) can be used for biological interactions in WikiPathways. We applied shape expressions to the harmonized interaction types and inform the users and curators what to expect as the shape of the interaction. Along with continuous curation of WikiPathways, we can improve the pathway drawings and their conversion on a regular basis.

The two examples show how the basic interaction information can be used, when semantically represented. This chapter also shows how interactions can be harmonized, but it cannot harmonize things that are not drawn consistently. This brings the discussion to harmonization of the interaction types. In this harmonizing step the interaction types, regardless of the drawing schema used to create the interaction, are given a common type. The examples used show how it is possible to use the modeled interactions from WikiPathways in order to answer different questions where connections between elements of a pathway are important. In specific instances we see the importance of directions of the interactions involved. This shows causality and is an important concept in understanding of normal bodily functions and also disease progression. The idea that changes in one portion can disrupt an organism system is an important concept in network biology and this is why it is necessary to model the interactions for a pathway resource.

7.3 WikiPathways in the Open PHACTS Discovery Platform

With the notion that we can represent biological pathways in semantic web formats, that we can harmonize the meaning of interactions, and that this can be used to study biological questions, the next question is if this can be used in drug discovery. Therefore, the WikiPathways RDF and the connectivity data was incorporated into the Open PHACTS Discovery Platform (OPDP) (*15*). Chapter 4 described the application of Chapters 2 and 3 in drug discovery. The OPDP already had information about small molecules and other drugs via ChEMBL (*16*), as well as information about their intended targets as well as disease information via UniProt (*17*). The addition of WikiPathways data added pathway and connectivity information for these targets. Because the OPDP does not provide direct access to the SPARQL endpoint, but via a REST-like Application Programming Interface(API), new API calls had to be developed and incorporated into the platform to query the pathway interaction data to return information about how targets are connected with one another.

This allowed Open PHACTS to become a uniform platform with information about drugs and targets and now the associated connections between targets allows more understanding of biology. Knowing how drug targets are connected with one another as well as information about their influence adds versatility to the platform. This endeavor saw the successful integration of WikiPathways semantic data into the Open PHACTS platform and creation of queries to navigate pathway diagrams. The adding of the queries associated with connectivity and directional influence was a powerful addition to the project and allowed for researchers to leverage this information in their own work in the realm of pharmacology and drug research. This additional information and being able to query it via the Open PHACTS API was an addition to the project with applications in drug repurposing and repositioning (*18*).

The API calls that were added to the Open PHACTS Discovery Platform may be straightforward, but using workflow tools, these tools can be used in more complicated workflows to leverage data from WikiPathways as well as from compound and gene information found in the rest of the platform. Tools like KNIME and Pipeline Pilot are examples of such workflow tools that work with the Open PHACTS Platform to create these new flows (*19–21*). There are also client libraries for JavaScript, R, and Neo4j that can access the OPDP. For example, a directional call can be made to find an alternative target and then a call to the OPDP for chemistry information about this new target for alternate active compounds. These workflows can make it possible to understand more context for pharmacology information.

It is important to remember that this is proof of concept that the addition of interaction queries to the Open PHACTS Drug Discovery Platform enriches the platform to make the platform able to answer more biology in one place. One does not have to go to many different resources to do research and data can be accessed from a single API. The API is focused on pharmacology but it also contains pathway, metabolite, drug data, gene, and gene disease associations which allows other areas of biology to be studied too. In this paper we proposed to answer pre-defined questions about biology using the Open PHACTS Discovery Platform (*15*). A platform that links drug information to gene information, pathways, and gene/protein connectivity information allows the user to identify alternative drug targets that are useful in drug repurposing and drug repositioning, as discussed elsewhere in the thesis. The continued update of platform data along with the platforms continued availability is required to keep the platform up to date and thus identifies a limitation of the platform along with most data backed platforms. The calls added to the platform allowed us to perform these alternative targets.

7.4 Subnetwork Analysis

With a large network of biological interactions now readily available, it is important to explore how these large network can be used efficiently. A large network describing everything about the biology of an organism is too complex to grasp. Therefore, we had to explore how subnetworks can be discovered from the large network. The active subnetworks chapter explores how to identify networks from a larger network in order to identify elements and connections in a rare disease and specifically Rett Syndrome. These subnetworks allow us to identify smaller active networks from datasets that are available and relate them to the condition being observed. Although Rett Syndrome was used for the study the same approach and principles can be applied to other diseases or any process that is less studied. In areas that are less studied with regards to the connections between elements, this approach proposes a network of elements and connections that are predicted to be active. This is done by constructing a larger network that represents the larger human network by combining common elements of pathways to link pathway diagrams together in order to create a large human network. New connections being added reflects our increasing understanding of how the system works together.

There are two important elements in order to make this work, a larger network that is reflective of the system that one is trying to observe and an appropriate dataset the defines what aspects of the larger network are worth zooming in on. We constructed our network for human pathways from WikiPathways. We then identified a suitable transcriptomics dataset specific for Rett Syndrome that was publically available on GEO (*22*). The dataset specifies which genes in the large network should be part of the subnetwork to be created. That is, it is possible to use the information from the dataset to identify the portions of the network that have changes in their expression and construct a network from this. Pathway analysis was also performed and it can be seen that from the dataset, the elements were distributed across multiple pathways. In the case of the active subnetworks, we have a single network that represents the expression data from our dataset, instead of the biology captured in individual drawings.

These approaches take advantage of all the interactions in the network, giving context to just a list of differentially expressed genes. This can be especially useful in the case of rare diseases where we have transcription data but not much information about how the transcript elements work together to have an influence on the system. In the case of rare diseases, they are often less well studied so this approach can be a boon to increase their understanding of how elements are connected when less is known about connectivity. This is a quick way to propose a disease-relevant network when research resources are limited.

This was the first time that the entire WikiPathways knowledge was used to construct a complete human network based upon the human pathway diagram found in WikiPathways. The construction of this human network enables more coverage than individual pathways can convey. This means that any dataset can be used against the network to identify new subnetworks that help describe the disease or process being studied. The complete network from WikiPathways is too large to usefully visualize in Cytoscape but the subnetworks that are more specific to the process being studied are more manageable (*23*). These subnetworks have the advantage of not being bound by artificial human boundaries of a pathway drawing that need to be drawn smaller for humans to more easily understand. Overlap between pathways helps us construct this larger human network. The subnetworks allow us to study processes and diseases that are oftentimes not specifically drawn in pathway diagrams. This has advantages of being able to perform analysis of more poorly understood diseases or processes. We can apply the knowledge of the larger human network to identify active subnetworks of more poorly described processes to pinpoint the relevant components and interactions in these processes. This has big implications in the area of rare diseases where less knowledge is often known. A researcher can apply this approach by using connectivity data described in Chapters 2 and 3 to other datasets on GEO to learn more about human biology and further their own research.

The subnetwork approach presented in Chapter 4, however, has one important limitation: the new network was constructed with regard to only how the elements of the network are connected with one another and it does not take into account the directions of the interactions connecting the elements of the network. Without directions as a part of the network, it is not possible to imply causality within the network, but adding directions allows the examination of which elements are responsible for elements downstream of it. Adding directions to the network is the next step needed to complete these networks. This also furthers the research to add more understanding of how the elements of these networks influence each other.

7.5 Drug Synergy

Finally, we wanted to know if an interaction network can be used for drug synergy predictions. The idea was to take pairs of cancer drug combinations and predict which drug pairs would be predicted to be synergistic without the aid of training data (*24*). We wanted to be able to use biological information that we know about pathways in order to predict the effect of drug combinations. This involves using both the connectivity information and directional information about interactions between drug targets to identify suitable drugs that target these genetic elements. Finding combinations that were connected via common paths through the pathway was the objective and then determine the synergy of the combination.

In order to tackle this problem, we used a previously known mathematical model, the Loewe Additivity Model (*25*), for predicting whether two drugs would be synergistic, but applied it using different parameters in order to predict which combinations are synergistic to one another. The different parameters used were expanding using two drug targets sharing a common pathway, connections across pathways, and the two targets sharing a common path of interactions with each other. The major assumption for the Loewe Additivity Model is that in order to predict if two drugs in combination are synergistic, the two drugs must share the same target. We expanded this model in several different scenarios in order to use the model when two drugs are not sharing the same target. The first scenario being that since a pathway diagram typically describes a biological function of an organism, then two potential targets of the drugs should be involved in similar functions, so in this scenario, drug combinations were created for the case that two targets share the same pathway. This includes neither connectivity nor directionality. The next scenario was to assume if two drug targets were found in different pathways but connected via a common interaction between them, then they are adjacent to each other and these combinations of drugs were calculated if they were synergistic or not. The final scenario using connections between targets that they have in common. If target A is connected to protein B and B is connected to target C, then we propose that combination targets A and C might be closely enough related in process to each other to apply the Loewe Additivity Model to predict whether the combinations are synergistic.

The objective being that if we can take things we do know about synergy and pathway connections, can we predict synergy without the aid of a training dataset. Since we have no background data about which combinations were synergistic or not to train a model we had to use prior knowledge to propose a method that could still predict a combination's synergy. We were then able to use knowledge of biological connectivity to propose predicted synergistic drug combinations to be scored by the organizers for their accuracy. This leveraging of data from sources outside the challenge data that we already have allows us to make predictions on synergy even if we would normally require more data to try and make these predictions about drug synergy of individual drug combinations.

7.6 Discussion

The common theme across all the papers is how information about how biological entities in a pathway are connected to one another help computational models gain more understanding of biology. Particularly, this thesis studies how connectivity and directional information of pathway elements support gaining insights to answer biological questions and expand our understanding of biology. This connectivity information has applications to better understanding of various topics in biology like identifying disease networks within a larger network. This connectivity information can be useful in the area of pharmacology when looking for alternative targets for either repositioning or repurposing of current drugs. Finally, in the case of predicting drug synergy for combinations of drugs, the connection and directional information about how two targets are connected is useful when we lack adequate outcome data to predict a combination's synergistic nature or not. The common thread across all outcomes being that in biology we have a system of elements that work together in order for an organism to function normally. If we see disruptions in elements of the system, we can cause changes to happen elsewhere in the system. In the case of disease, a process is not acting normally and causes changes elsewhere. For pharmacology, we use small molecules in order to bring these functions back to normal states that we can use to treat patients with abnormal functions. This thesis has shown how the system elements work together in Chapters 2 and 3, in Chapters 4 and 6 we see the applications for pharmacology, and Chapter 5 we have seen the application of the interactions in rare disease networks.

Being able to analyze an entire biological process has the advantage of being able to observe effects outside of the immediate pathway element or relationship that we normally study. Connectivity helps with the understanding of detailed biology, often including cellular and tissue location, causality, activation, and more. Also having information of directional influences enriches this understanding even further. Being able to understand the fundamental causality of influence is a powerful tool in biological research. It has implications in the understanding and progression of disease as well as with the treatment processes in pharmacology. It also can be applied in areas like toxicology to study how chemicals outside the system might alter the expression of the organism and the cascade of effects that can be observed through that process. The papers described in the previous chapters are applications of how this research can be used in order to advance our very understanding of biology itself. This was shown in multiple research lines and puts on full display the power of this systematic approach to understanding biology. This is not the limits of the research that can be applied using these approaches, making it potentially even more powerful for understanding other areas of biology.

Connectivity and directional information giving a more specific understanding of how the system works together by making pathway information readily available and queryable to the scientific community is an essential concept in biology and why this research is worth pursuing. This thesis demonstrates how to use this connectivity information in several fields of biology. This work was done in order to advance our understanding of how biological processes work and work together to help an organism survive. The thesis advanced the field by making connectivity and directional interactions available to the community for pathway networks and showing how they can be used in the areas of pharmacology, rare diseases, and systems biology. In itself this is not new: approaches like the SBML do this too (*26*). WikiPathways is also flexible in its drawing and pathway drawing do not always have the necessary components to make a formal SBML drawing possible. This work shows, however, that with semantic web approaches we can harmonize interaction knowledge in biological diagrams in different formats, including MIM and SBGN (see Chapter 3). Connection and directional information is a fundamental concept in biology and understanding how they work together has applications in several fields of living organisms. The thesis work has wide ranging implications in biological research. The work makes it possible to use connectivity data to repurpose drugs through platforms like Open PHACTS, to propose drug combinations that have the potential to be synergistic, and propose subnetworks of the larger human network to suggest active subnetworks in rare diseases. The thesis work from the thesis chapters presented show just how useful this work can be in advancing scientific study in biology.

While the thesis was able to show how and why this work is important, there still remain some areas for development and some limitations. While the connectivity and directional information is useful in its current form, it does depend on continual development and curation of pathway sources. This was demonstrated in Chapter 3 where we show that interaction data in WikiPathways is incomplete. A platform like Open PHACTS is helpful but it requires continual updates to maintain current and up to date connections that reflect our latest understanding of biology. This is painfully clear when you realize that the original ODPD has been offline for some time now. Predicting drug synergies using directional information could be improved with a more robust mathematical model to calculate the synergies of the combinations. The inclusion of directional information to the active subnetwork connections would give a more accurate representation of the model that we are trying to recreate. One can image how federated approaches can extend the semantic pathway knowledge with detailed information about enzymatic reactions, ligand-protein binding, and protein-protein interactions. These outweigh the limitations involved with the work presented here, and makes for interesting future development to enhance our understanding of these biological networks further.

References

1. M. Kutmon *et al.* WikiPathways: capturing the full diversity of pathway knowledge. *Nucleic Acids Research*. 2016. **44** (D1): D488–D494.
2. L. Jin *et al.* Pathway-based Analysis Tools for Complex Diseases: A Review. *Genomics, Proteomics & Bioinformatics*. 2014. **12** (5): 210–220.
3. Z. Yan, Z. kui, Z. Ping. Reviews and prospectives of signaling pathway analysis in idiopathic pulmonary fibrosis. *Autoimmunity Reviews*. 2014. **13** (10): 1020–1025.
4. S. Zhao, R. Iyengar. Systems Pharmacology: Network Analysis to Identify Multiscale Mechanisms of Drug Action. *Annual Review of Pharmacology and Toxicology*. 2012. **52** (1): 505–521.
5. M. R. Karimi, A. H. Karimi, S. Abolmaali, M. Sadeghi, U. Schmitz. Prospects and challenges of cancer systems medicine: from genes to disease networks. 2021. : .
6. I. Tassi, J. Klesney-Tait, M. Colonna. Dissecting natural killer cell activation pathways through analysis of genetic mutations in human and mouse. *Immunological Reviews*. 2006. **214** (1): 92–105.
7. M. Kutmon *et al.* PathVisio 3: An Extendable Pathway Analysis Toolbox. *PLOS Computational Biology*. 2015. **11** (2): e1004085.
8. D. R. Zerbino *et al.* Ensembl 2018. *Nucleic Acids Research*. 2017. **46** (D1): D754–D761.
9. S. M. Wimalaratne *et al.* Uniform resolution of compact identifiers for biomedical data. *Scientific Data*. 2018. **5** (1): 180029.

10. C. R. Gault, L. M. Obeid, Y. A. Hannun. An overview of sphingolipid metabolism: from synthesis to breakdown. *Advances in experimental medicine and biology*. 2010. **688**: 1–23.

11. S. M. Kyle, P. K. Saha, H. M. Brown, L. C. Chan, M. J. Justice. MeCP2 co-ordinates liver lipid metabolism with the NCoR1/HDAC3 corepressor complex. *Human Molecular Genetics*. 2016. : ddw156.

12. J. Guy, H. Cheval, J. Selfridge, A. Bird. The Role of MeCP2 in the Brain. *Annual Review of Cell and Developmental Biology*. 2011. **27** (1): 631–652.

13. M. Wan *et al.* Rett Syndrome and Beyond: Recurrent Spontaneous and Familial MECP2 Mutations at CpG Hotspots. *The American Journal of Human Genetics*. 1999. **65** (6): 1520–1529.

14. A. Waagmeester *et al.* A protocol for adding knowledge to Wikidata: aligning resources on human coronaviruses. *BMC Biology*. 2021. **19** (1): 12.

15. K. Azzaoui *et al.* Scientific competency questions as the basis for semantically enriched open pharmacological space development. *Drug Discovery Today*. 2013. **18** (17): 843–852.

16. A. Gaulton *et al.* The ChEMBL database in 2017. *Nucleic Acids Research*. 2016. **45** (D1): D945–D954.

17. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*. 2018. **47** (D1): D506–D515.

18. S. Ekins, A. J. Williams, M. D. Krasowski, J. S. Freundlich. In silico repositioning of approved drugs for rare and neglected diseases. 2011. **16** (7-8): 298–310.

19. W. A. Warr. Scientific workflow systems: Pipeline Pilot and KNIME. *Journal of Computer-Aided Molecular Design*. 2012. **26** (7): 801–804.

20. M. R. Berthold *et al.* "KNIME: The Konstanz Information Miner". Paper presented at: Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007); ; : Springer; 2007. .

21. BIOVIA - Dassault Systèmes, *Pipeline pilot*, https://www.3ds.com/products-services/biovia/products/data-science/pipeline-pilot/.

22. T. Barrett *et al.* NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research*. 2012. **41** (D1): D991–D995.

23. P. Shannon. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*. 2003. **13** (11): 2498–2504.

24. M. P. Menden *et al.* Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. *Nature Communications*. 2019. **10** (1): 2674.

25. W. E. Delaney, H. Yang, M. D. Miller, C. S. Gibbs, S. Xiong. Combinations of Adefovir with Nucleoside Analogs Produce Additive Antiviral Effects against Hepatitis B Virus In Vitro. *Antimicrobial Agents and Chemotherapy*. 2004. **48** (10): 3702–3710.

26. I. Balaur *et al.* Recon2Neo4j: applying graph database technologies for managing comprehensive genome-scale networks. *Bioinformatics*. 2016. **33** (7): btw731.

Summary

The conclusion of this thesis is that we model and describe the biological interactions of a resource such as WikiPathways as needed. The implications are that we can further our understanding of biology by tuning the amount of detail we need. This is accomplished by knowing how the portions of a pathway diagram interact with each other and what sort of influence can be observed. Understanding not just that they interact with each other but how they interact with each other is a significant development to the platform. The first two chapters of this thesis describe how the WikiPathways knowledgebase is modelled for both data nodes and edges. This addition enriches WikiPathways as a platform for studying system effects of changes in gene product expression.

This had influences outside of the platform. Even though WikiPathways is primarily used for pathway analysis of changes in gene expression or in metabolomics studies, WikiPathways can also be used for whole network analysis of an organism such as *Homo sapiens* as seen in the chapter dealing with identifying active subnetworks in Rett Syndrome. This demonstrates how WikiPathways data can be used to identify proposed subnetworks using an appropriate dataset. This can be used to study disease subnetworks and is particularly practical in the case of rare diseases where less is known about the underlying mechanisms of connectivity to each other.

The chapter relating to cancer biology and the Open PHACTS Discovery Platform are clear indications of how WikiPathways can be used in the realms of pharmacology as well as general toxicity. Open PHACTS is a platform designed with pharmacology in mind and as such has an interest in compound interactions. The inclusion of the WikiPathways knowledgebase with appropriate queries to the Open PHACTS application programming interface (API) meant that a new class of research

questions were able to be answered with the connection information for potential targets. Research questions relating to drug repurposing by finding alternate targets are now possible.

In the case of cancer biology, we were tasked with identifying potential drug combinations that we would predict to be synergistic. Using information provided by the DREAM challenge coordinators relating to drug dosage concentrations to achieve cell death for cancerous cells and the interaction information from WikiPathways we were able to propose several lists of drug pairs that could be potentially synergistic. These lists were dependent upon our understanding of how drug targets are connected with each other and how they affect similar processes. Constructing pairs of drugs that affect similar processes in this way takes advantage of our collective previous knowledge in a new way to generate lists of potentially synergistic drug combinations.

The thesis and its papers have demonstrated how valuable connectivity information can be to further biological understanding of how systems work together. The information itself is a valuable commodity to be used in scientific analysis. The description of how interactions are modeled within the WikiPathways ecosystem makes the platform available for more users to understand and deploy in their own solutions. Several examples of how this information can be leveraged to answer scientific questions have been explored in areas such as pharmacology, rare diseases, and cancer biology have been explored throughout this thesis and provide a solid foundation for further work to be done in these areas or in system biology studies in general.

The inclusion of the work done and its advanced understanding allows the answering of more questions related to biological processes and is an advancement of studies in biology. Using the techniques and work done for this thesis, it is possible to study biology from the perspective of how biological systems work together or from the perspective of pharmacology, toxicity and systems biology in general.

Impact

The work described in this thesis regarding how interactions within pathways define biological processes and the larger biological system as a whole has impacts that can be seen in other areas outside systems biology. This can be seen in areas such as pharmacology. Published papers pertaining to WikiPathways have many citations with many discussing their desire to integrate WikiPathways data and network information into other platforms. The paper pertaining to WikiPathways connections practical effects can be seen in areas like drug synergy where connectivity of targets is important. The work done with interactions and WikiPathways was incorporated into a larger IMI project called the Open PHACTS Discovery Platform and linked WikiPathways connectivity information with drug chemistry and gene product data. The platform made the data available via a publically available API that researchers could use in drug discovery and general pharmacological research. These large European IMI projects also affect future work in areas like medicine as a whole. Thus the work is anticipated to reach a wide audience outside of systems biology.

The work was also put to use in one of the DREAM challenges, which proposes to use community expertise and competition to further understanding in a particular area. In this case the work was used to address the area of drug synergy in cancer research. This work has an impact in medicine, pharmacology, cancer research, and toxicology. It has a very wide audience that can use my work to advance their understanding of these topics.

The wider implication of this work means that the description of interactions from WikiPathways can assist biologists in their understanding of how their genes and their proteins influence or be influenced by other elements in the system. This goes to the idea of causality and how changes in one portion of the system have effects that can be observed in other portions of the pathway. Processes being described need not only the DataNode elements also require descriptions of the edges that connect the nodes. These edges can be directed or undirected as well as have information about the type of interaction that it is. Explicit semantics that are formalized and contain links to the meaning in the pathway is currently being used by the WikiPathways project to automatically find inconsistencies and pathways that have potential need of biological curation.

This greater understanding of how pathway portions are connected with each other has consequences beyond studies in systems biology. It has wider applications for the field of biology as a whole and allows biologists to know how their own genes and proteins that they study fit into a biological system. Understanding how their proteins influence or are influenced by other portions of the pathway gives Biologists important context for how their proteins affect biological processes.

The work advances our understanding of science in a more general sense. Principles from systems biology can be used in any network. Networks are becoming more important in science and technology. The principles of which are similar in both the case of biology and other types of networks. Network analysis approaches are the same regardless of the application. Data nodes connected by edges are how we analyze biological networks but this modelling and description is used throughout science and technology.

This type of work and thinking can be seen mirrored in many areas in our lives. Systems build upon each other to build up complexity. First we have individual atoms and molecules of elements that work together to form an organism. Then we have organisms that work together and we can see reflected in nature. This build up of complexity of systems forms the world that we know. Our world itself is part of an even larger system that forms solar systems and galaxies and beyond. We can even see networks that form between people that forms our social structures. This shows the pervasiveness of systems of things working together toward a larger end.

Acknowledgments

I would like to take this opportunity to thank all the people that have made this thesis work possible. I would not have been able to complete this work without all of your work and all of your contributions to my life. I start by thanking my promoter, Chris Evelo. You were always there to push me harder and encourage me to pursue research. You saw something in me that made this work a possibility. I also need to thank my co-promoter, Egon Wilighagen. We did not always see eye to eye on everything, but you were always driving me to be a better researcher and all your hard work with me has also made this work possible.

Next, I need to thank all the co-workers who helped make my work come to fruition. Martina Kutmon, you have been around since the beginning of my studies in Maastricht. You helped me in so many ways both personally and professionally. Lars Eijsen you were both a professional source that helped me advance my career but also a good friend when I needed a friendly ear. Fredierieke Ehrhart you were also a person I count as a friend and also thank you for your professional advice as well. Linda Reiswijk, you were someone I enjoyed working with very much and your perspective was always welcomed on projects.

I, of course, must thank my co-authors. I enjoyed working with all of you. Your insights helped direct our papers to just be better. Every one of you contributed and helped me in an important way and I cannot thank you enough for your assistance and support along the way. My fellow co-authors are Andra Waagmeester, Anders Riutta, Alexander Pico, Peter Woollard, Daniela Digles, Antonis Loizou, Stefan Senger, Denise Slenter, Leopold Curfs, Nuno Nunes, Kristina Hanspers, Anwesha Bohler, Jonathan Mélius, Andra Waagmeester, Sravanthi Sinha, Susan Coort, Elisa Cirillo, and Bart Smeets.

I also need to acknowledge my co-workers and office mates that helped me as well. Whether we worked directly with one another on manuscripts or not, our interactions were important to me and I could not have gotten through the rigors of the PhD works without your help. Susan Coort, Elisa Cirillo, Bart Smeets, Amadeo Muñoz-Garcia, Nuno Nunes, Jonathan Mélius, Mirella Kalafati, Marvin Martens, and Denise Slenter we did not always work directly with each other for many publications, but we shared common goals and you all gave me valuable insights.

Finally it is necessary to thank my family and friends. My parents, Cynthia and Stephen Miller, have always been there for me from the beginning right up through my PhD work. My grandparents, Miriam and Stanley Miller, were like a second set of parents that took care of me growing up and always encouraged me to pursue higher education. Jonathan Sheeley, Michael Bolger and I have been close friends from the early grade school years into the present. We have maintained our friendship whether in another city, another state, or another country.

There are countless other people that I have met and interacted with throughout my life and although I may not be close to you, I thank you for helping to shape me into the person that I am today. My friends, family, co-workers, promoters and neighbors, you have all touched me in a profound way and I am glad to have known all of you and am glad you helped me get to this point in my life. So in the most sincere manner possible, I thank you all.

Ryan A. Miller
Maastricht
2022-12-06

About the author

Ryan A. Miller was born on December 20, 1982 in Camp Hill, Pennsylvania, USA. He graduated from high school in June of 2001. He graduated with his bachelor's degree in Biotechnology from Harrisburg University of Science and Technology in May 2010. He then went on to study Bioinformatics at the University of the Sciences in Philadelphia, Pennsylvania. His master's project was overseen by Dr. Randy Zauhar, Professor of Chemistry and Biochemistry. He went on to receive his master's degree in Bioinformatics in May of 2013. Starting in January of 2015, Ryan began his PhD research in the Department of Bioinformatics - BiGCaT at Maastricht University. There he helped further the development of WikiPathways and researched pathway connections and directional elements under the supervision of Professor Chris Evelo and Assistant Professor Egon Willighagen.

Published work

Kutmon, M., Riutta, A., Nunes, N., Hanspers, K., Willighagen, E. L., Bohler, A., Mélius, J., Waagmeester, A., Sinha, S. R., **Miller, R.**, Coort, S. L., Cirillo, E., Smeets, B., Evelo, C. T., Pico, A. R. (2015). WikiPathways: capturing the full diversity of pathway knowledge. Nucleic Acids Research, 44(D1), D488–D494. https://doi.org/10.1093/nar/gkv1024

Waagmeester, A., Kutmon, M., Riutta, A., **Miller, R.**, Willighagen, E. L., Evelo, C. T., Pico, A. R. (2016). Using the Semantic Web for Rapid Integration of WikiPathways with Other Biological Online Data Resources. PLOS Computational Biology, 12(6), e1004989. https://doi.org/10.1371/journal.pcbi.1004989

Slenter, D. N., Kutmon, M., Hanspers, K., Riutta, A., Windsor, J., Nunes, N., Mélius, J., Cirillo, E., Coort, S. L., Digles, D., Ehrhart, F., Giesbertz, P., Kalafati, M., Martens, M., **Miller, R.**, Nishida, K., Rieswijk, L., Waagmeester, A., Eijssen, L. M. T., Evelo, C.T., Pico, A.R., Willighagen, E. L. (2017). WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. Nucleic Acids Research, 46(D1), D661–D667. https://doi.org/10.1093/nar/gkx1064

Miller, R. A., Woollard, P., Willighagen, E. L., Digles, D., Kutmon, M., Loizou, A., Waagmeester, A., Senger, S., Evelo, C. T. (2018). Explicit interaction information from WikiPathways in RDF facilitates drug discovery in the Open PHACTS Discovery Platform. F1000Research, 7, 75. https://doi.org/10.12688/f1000research.13197.2

Miller, R. A., Ehrhart, F., Eijssen, L. M. T., Slenter, D. N., Curfs, L. M. G., Evelo, C. T., Willighagen, E. L., Kutmon, M. (2019). Beyond Pathway Analysis: Identification of Active Subnetworks in Rett Syndrome. Frontiers in Genetics, 10. https://doi.org/10.3389/fgene.2019.00059

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. (2020). WikiPathways: connecting communities. Nucleic Acids Research, 49(D1), D613–D621. https://doi.org/10.1093/nar/gkaa1024

Miller, R. A., Kutmon, M., Bohler, A., Waagmeester, A., Evelo, C. T., Willighagen, E. L. (2022). Understanding signaling and metabolic paths using semantified and harmonized information about biological interactions. PLoS ONE, 17(4), e0263057. https://doi.org/10.1371/journal.pone.0263057

Miller, R. A., Muraru S., Belén Malpartida A., Low J., Pelicano de Almeida M., van Engelen B., Godynyuk E., Schmitz-Abecassis, B., Willighagen E., Evelo C., Rieswijk L. (2022). An approach to the proposal of drug combination for cancer therapy using a pathway data connectivity approach. ChemRxiv. Cambridge: Cambridge Open Engage. https://doi.org/10.26434/chemrxiv-2022-0n8fp

Menden M.P, Wang D., Mason M.J., Szalai B., Bulusu K.C., Guan Y., Yu T., Kang J., Jeon J., Wolfinger R., Nguyen T., Zaslavskiy M., AstraZeneca-Sanger Drug Combination DREAM Consortium, **Miller, R.A.**, Jang I.S., Ghazoui Z., Ahsen M.E., Vogel R., Neto E.C., Norman T., Tang E.K.Y., Garnett M.J., Di Veroli G.Y., Fawell S., Stolovitzky G., Guinney J., Dry J.R., Saez-Rodriguez J. (2019). Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. Nature Communications, 10. https://doi.org/10.1038/s41467-019-09799-2