

# Computational Analysis of Pathophysiological Mechanisms Based on Pathway Modeling

Kumulative Dissertation

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

vorgelegt von

DANIEL DOMINGO FERNÁNDEZ

aus Huéscar, Spanien

Bonn, 2019



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

1. Gutachter: Univ.-Prof. Dr. rer. nat. Martin Hofmann-Apitius
  2. Gutachter: Univ.-Prof. Dr. rer. nat. Andreas Weber
- Tag der Promotion: December 9, 2019  
Erscheinungsjahr: 2019



*"Ever tried. Ever failed. No matter. Try Again. Fail again. Fail better."*

*- Samuel Beckett*



# Abstract

The advent of the big data era poses major challenges to the biomedical domain. First, it is necessary to adopt strategies that integrate and link the heterogeneous resources that contain multiscale and multimodal data in order to fill the existing knowledge gaps. Further, there is a need for developing methods designed not only to interrogate the data but also to interpret and decode the complex world of biology.

In this work, we address the two aforementioned challenges in the domain of pathway knowledge. This thesis presents two ecosystems devised to harmonize and consolidate knowledge from disparate pathway databases, ultimately providing a holistic view of the pathway landscape. Leveraging this integrative effort, we designed a benchmarking study that demonstrates significant impact of database selection in functional enrichment methods and prediction modeling. The results of this work advocate for integrative approaches since our unifying schema has been shown to yield more robust and interpretable results than individual databases and to improve the predictability in modeling tasks. Tangential to these pathway-driven approaches, this work also presents two frameworks devised to identify mechanisms and biomarkers in the neurodegenerative and psychiatric field. The first resource, NeuroMMSig, is the largest inventory of candidate mechanisms for Alzheimer's and Parkinson's disease. This manually-curated collection of over 200 computable mechanistic networks emerged as a novel knowledge-based paradigm by laying the ground for the first draft of a mechanism-based taxonomy in both conditions. The second resource, PTSDDB, is a database cataloging biomarker information in the context of post-traumatic stress disorder that opens the door for a future systematic meta-analysis of results reported in literature. Finally, we conclude the thesis with a novel approach that bridges the gap between mechanistic knowledge and patient-level data, paving the way for a mechanism-based stratification of dementia patients.

In summary, this thesis presents novel methodologies for the integration of pathway knowledge. In addition, it introduces new resources and strategies in the context of neurodegenerative and psychiatric disorders. These advances have numerous applications in translational research, ranging from drug discovery to patient stratification.





# Acknowledgment

This thesis is the result of years of effort and I want to personally thank those who helped with this work.

To my parents, friends, and to my girlfriend, thank you for always allowing me to chase my dreams. Without your continued support throughout the years, this thesis would not have been possible.

To my supervisor, Prof. Dr. Martin Hofmann-Apitius, thank you for the opportunities and challenges you have given to me and all the support and trust I had as a scientist and as a person since I started as a young student in the department. Your teachings, first as professor in B-IT and then as *boss* have not only instilled in me passion about science but also to look at the world from other perspectives. Moreover, to Prof. Dr. Andreas Weber, thank you for your help and for acceding to be the second reviewer of this thesis. Finally, to Prof. Dr. Jürgen Bajorath and Prof. Dr. Diana Imhof, thank you for being in my defense committee.

To all my SCAI-BIO colleagues, a heartfelt thank you! Not only have you, as a team, helped me to reach this goal, but also to have the strength to wake up every morning with a smile for doing what I like where and with the people I wanted. You really made the *PhD road* flatter than my beloved commuting climb to the Fraunhofer Campus. A special thanks goes to my first supervisor Dr. Alpha Tom Kodamullil and to Dr. Charles Tapley Hoyt. The next thank you goes to all the "students" who contributed to this work, particularly to Sarah Mubeen and Josep Marín-Llaó. Furthermore, I would like to thank the *PhD gang* for all the support and great working atmosphere we have. Finally, I cannot finish these acknowledgments without thanking my *sporty* seniors and core of the department: Alina Enns (*meine kleine Schwester und Fitness Kollegin*), Stephan Springstube (*der Pate und Tischtennis Master*), Meike Knieps (*die Mama und Läuferin*), and Prof. Dr. Juliane Fluck (*Tennis Königin von Fraunhofer*). Thank you for being there with a smile when I needed you.



# Publications

## Thesis publications

<sup>†</sup> Contributed Equally

- **Daniel Domingo-Fernández**, Charles Tapley Hoyt, Carlos Bobis Álvarez, Josep Marín-Llaó, and Martin Hofmann-Apitius. "ComPath: An ecosystem for exploring, analyzing, and curating pathway databases". *npj Systems Biology and Applications*, Volume 4, 43, (2018).  
<https://doi.org/10.1038/s41540-018-0078-8>
- **Daniel Domingo-Fernández**, Sarah Mubeen, Josep Marín-Llaó, Charles Tapley Hoyt, and Martin Hofmann-Apitius. "PathMe: Merging and exploring mechanistic pathway knowledge". *BMC Bioinformatics*, Volume 20, Article number 243 (2019).  
<https://doi.org/10.1186/s12859-019-2863-9>
- Sarah Mubeen, Charles Tapley Hoyt, André Gemünd, Martin Hofmann-Apitius, Holger Fröhlich, and **Daniel Domingo-Fernández**. "The impact of pathway database choice on statistical enrichment analysis and predictive modeling". *Frontiers in Genetics*, 10:1203 (2019).  
<https://doi.org/10.3389/fgene.2019.01203>
- **Daniel Domingo-Fernández**, Alpha Tom Kodamullil, Anandhi Iyappan, Mufassra Naz, Mohammad Asif Emon, Tamara Raschka, Reagon Karki, Stephan Springstube, Christian Ebeling, and Martin Hofmann-Apitius. "Multimodal Mechanistic Signatures for Neurodegenerative Diseases (NeuroMMSig): a web server for mechanism enrichment". *Bioinformatics*, Volume 33, Issue 22, Pages 3679–3681 (2017).  
<https://doi.org/10.1093/bioinformatics/btx399>
- **Daniel Domingo-Fernández**<sup>†</sup>, Allison Provost<sup>†</sup>, Alpha Tom Kodamullil<sup>†</sup>, Josep Marín-Llaó, Heather Lasseter, Kristophe Diaz, Lee Lancashire, Martin Hofmann-Apitius, and Magali Haas. "PTSD Biomarker Database: deep

dive meta-database for PTSD biomarkers, visualizations, and analysis tools". *Database: The Journal of Biological Databases and Curation*, Volume 2019, baz081 (2019).

<https://doi.org/10.1093/database/baz081>

- Shashank Khanna<sup>†</sup>, **Daniel Domingo-Fernández<sup>†</sup>**, Anandhi Iyappan, Mohammad Asif Emon, Martin Hofmann-Apitius, and Holger Fröhlich. "Using multi-Scale Genetic, neuroimaging and clinical data for predicting Alzheimer's disease and reconstruction of relevant biological mechanisms". *Scientific Reports*, Volume 8, Article number: 11173 (2018).

<https://doi.org/10.1038/s41598-018-29433-3>

## Other publications

<sup>†</sup> Contributed Equally

- Charles Tapley Hoyt<sup>†</sup>, **Daniel Domingo-Fernández<sup>†</sup>**, Nora Balzer, Anka Gueldenpfennig, and Martin Hofmann-Apitius. "A systematic approach for identifying shared mechanisms in epilepsy and its comorbidities". *Database: The Journal of Biological Databases and Curation*, Volume 2018, bay050 (2018).  
<https://doi.org/10.1093/database/bay050>
- Farah Humayun<sup>†</sup>, **Daniel Domingo-Fernández<sup>†</sup>**, Ajay Abisheck Paul George, Marie-Thérèse Hopp, Benjamin F. Syllwasschy, Milena S. Detzel, Charles Tapley Hoyt, Martin Hofmann-Apitius, and Diana Imhof. "A computational approach for mapping heme biology in the context of hemolytic disorders". *bioRxiv*, 804906 (2019).  
<https://doi.org/10.1101/804906>
- Charles Tapley Hoyt, **Daniel Domingo-Fernández**, and Martin Hofmann-Apitius. "BEL Commons: an environment for exploration and analysis of networks encoded in Biological Expression Language". *Database: The Journal of Biological Databases and Curation*, Volume 2018, bay126 (2018).  
<https://doi.org/10.1093/database/bay126>
- Mehdi Ali, Charles Tapley Hoyt, **Daniel Domingo-Fernández**, Jens Lehmann, and Hajira Jabeen. "BioKEEN: A library for learning and evaluating biological knowledge graph embeddings". *Bioinformatics*, btz117, (2019).

<https://doi.org/10.1093/bioinformatics/btz117>

- Charles Tapley Hoyt, **Daniel Domingo-Fernández**, Rana Aldisi, Lingling Xu, Kristian Kolpeja, Sandra Spalek, Esther Wollert, John Bachman, Benjamin Gyori, Patrick Greene, and Martin Hofmann-Apitius. "Re-curation and rational enrichment of knowledge graphs in Biological Expression Language". *Database: The Journal of Biological Databases and Curation*, Volume 2019, baz068 (2019).

<https://doi.org/10.1093/database/baz068>

- Charles Tapley Hoyt, **Daniel Domingo-Fernández**, Sarah Mubeen, Josep Marín-Llaó, Andrej Konotopez, Christian Ebeling, Colin Birkenbihl, Özlem Muslu, Bradley English, Simon Müller, Mauricio Pio de Lacerda, Mehdi Ali, Scott Colby, Dénes Türei, Nicolás Palacio-Escat, Martin Hofmann-Apitius. "Integration of structured biological data sources using Biological Expression Language". *bioRxiv*, 631812 (2019).

<https://doi.org/10.1101/631812>

- Eduardo Brito, Bogdan Georgiev, **Daniel Domingo-Fernández**, Charles Tapley Hoyt, and Christian Bauckhage. "RatVec: a general approach for low-dimensional distributed vector representations via domain-specific rational kernels". *Proceedings of LWDA-KDML* (2019).
- Mehdi Ali, Charles Tapley Hoyt, **Daniel Domingo-Fernández**, and Jens Lehmann. "Predicting Missing Links Using PyKEEN". *Proceedings of ISCW* (2019).
- Mohammad Asif Emon, **Daniel Domingo-Fernández**, Charles Tapley Hoyt, and Martin Hofmann-Apitius. "PS4DR: a multimodal workflow for identification and prioritization of drugs based on pathway signatures". *BMC Bioinformatics*, submitted, (2019).
- Sepehr Golriz Khatami, Christine Robinson, Colin Birkenbihl, **Daniel Domingo-Fernández**, Charles Tapley Hoyt, and Martin Hofmann-Apitius. "Challenges of Integrative Disease Modeling in Alzheimer's disease". *Frontiers in Molecular Biosciences*, submitted, (2019).



# Contents

|   |   |     |
|---|---|-----|
| 1 | Introduction  | 1   |
| 2 | ComPath: an ecosystem for exploring, analyzing, and curating mappings across pathway databases  | 25  |
| 3 | PathMe: Merging and exploring mechanistic pathway knowledge   | 37  |
| 4 | The impact of pathway database choice on statistical enrichment analysis and predictive modeling  | 53  |
| 5 | Multimodal Mechanistic Signatures for Neurodegenerative Diseases (NeuroMMSig): a web server for mechanism enrichment                              | 71  |
| 6 | PTSD Biomarker Database: deep dive meta-database for PTSD biomarkers, visualizations, and analysis tools  | 79  |
| 7 | Using multi-scale genetic, neuroimaging and clinical data for predicting Alzheimer's disease and reconstruction of relevant biological mechanisms | 91  |
| 8 | Conclusion and outlook  | 109 |





# Acronyms

**AD** Alzheimer's disease.

**ADNI** Alzheimer's Disease Neuroimaging Initiative.

**API** Application Programming Interface.

**BEL** Biological Expression Language.

**BioPAX** Biological Pathway Exchange.

**CAM** Causal Activity Models.

**CHEBI** Chemical Entities of Biological Interest.

**GO** Gene Ontology.

**HGNC** HUGO Gene Nomenclature Committee.

**HUPO** Human Proteome Organization.

**KAM** Knowledge Assembly Model.

**KEGG** Kyoto Encyclopedia of Genes and Genomes.

**KGML** KEGG Markup Language.

**MCI** Mild Cognitive Impairment.

**MSigDB** Molecular Signatures Database.

**NeuroMMSig** Multimodal Mechanistic Signatures for Neurodegenerative Diseases.

**OWL** Web Ontology Language.

**PD** Parkinson's disease.

**PPMI** Parkinson's Progression Markers Initiative.

**PSI-MI** Proteomics Standards Initiative - Molecular Interaction.

**PTSD** Post-traumatic stress disorder.

**PTSDDB** PTSD Biomarker Database.

**RDF** Resource Description Framework.

**SBGN** Systems Biology Graphical Notation.

**SBML** Systems Biology Markup Language.

**SIF** Simple Interaction Format.

**SNP** Single Nucleotide Polymorphism.

**SSRIs** Selective Serotonin Re-uptake Inhibitors.

**URI** Uniform Resource Identifier.

**XML** Extensible Markup Language.



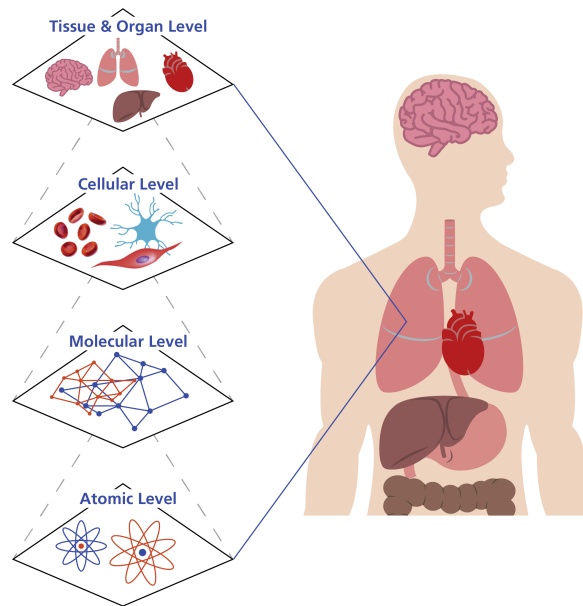


# 1 Introduction

The advent of big biodata, machine learning, and artificial intelligence brings high hopes that these state of the art technologies will lead to great advances in the biomedical field [1]. However, the structural and functional complexity of living organisms pose unique challenges for these approaches because they were not originally designed to interpret nor understand the mechanisms underlying biology. Living organisms are comprised of specialized and variable cellular and tissue structures comprised of molecules, which are essentially sets of atomic structures. Each of these can be considered biological levels of organization, that are not only regulated by their underlying changes but also by their interactions with other levels in this multi-scale hierarchy (**Figure 1**) [2]. As an illustration, an imbalance in the concentration of a given transcription factor can dysregulate the expression of multiple proteins, ultimately resulting in cell death and organ dysfunction. Furthermore, not only are particular species distinct, but each individual organism has a unique composition of different tissue and cellular types that are themselves constituted by millions of disparate biological entities. Hence, understanding biology involves revealing the causal interactions occurring between these entities, both in each of the mentioned scales and across them [3, 4].

## 1.1 Pathways: the functional units of biology

Because cells are the basic structural and functional units of living organisms, studying the interactions occurring at this level is essential to enhance our understanding of biology. However, though every cell in an individual organism typically contains identical genetic information, the context in which they reside



**Figure 1:** Multi-scale biology of a given organism in a bottom-up approach. Any biological organism can be subdivided in multiple scales depending on the level of granularity we want to study its structure or function. Here, this organizational representation of biology is depicted in terms of distinct scales of increasing complexity, from the atomic level to the tissue and the organ level. The growing need to understand the interactions across these different scales was what gave rise to the field of systems biology. This interdisciplinary field of study attempts to understand biology by modeling and analyzing these complex interactions using computational and mathematical methods.

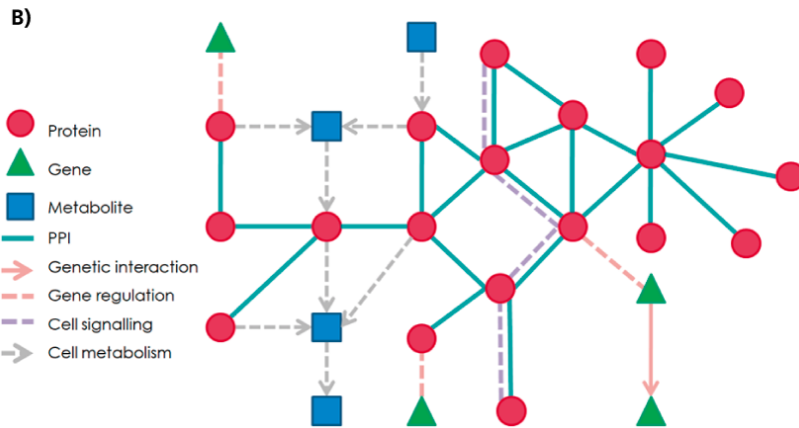
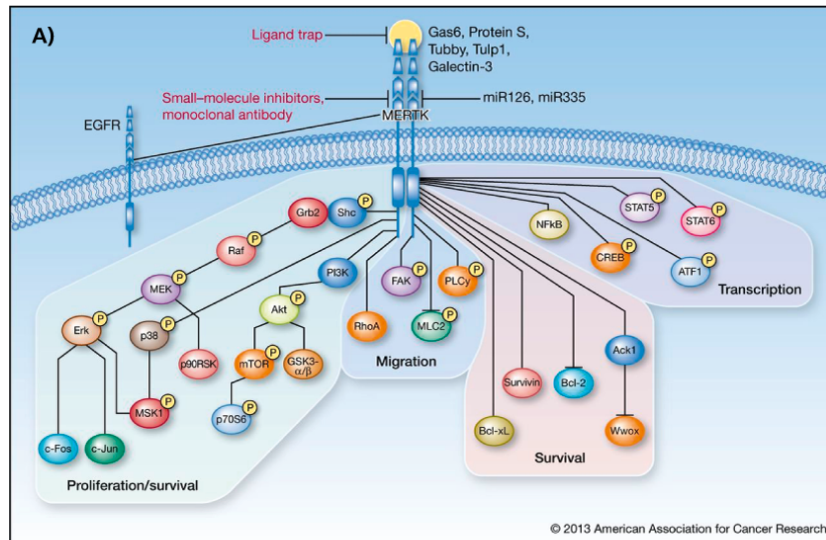
makes them well adapted for disparate specialized tasks [5]. In order to help us in deconvoluting the numerous processes that take place on both cellular and sub-cellular levels, humans conceived the notion of a pathway, which corresponds to a series of molecular interactions that leads to a particular event. This concept facilitates the representation, formalization, and interpretation of biological events by abstracting these series of interactions from a vast and complicated biological universe [6]. In other words, cataloging biological knowledge into pathways reduces complexity from all possible interacting molecular entities to sets of well-studied and validated functional relationships between entities that culminate in specific biological processes.

Pathways are usually represented as networks or mathematical models<sup>1</sup>. However, simplifying biology into any human-fabricated representation inevitably results in a loss of information, such as spatio-temporal information, or even ignores certain biological entity types altogether [7]. Nonetheless, a network abstraction can facilitate pathway visualization and interpretation on account of its concordance with biological systems: nodes correspond to molecular entities (e.g., genes, proteins, chemicals, etc.) and edges to types of interactions occurring between them (e.g., inhibition, phosphorylation, etc.) (**Figure 2**). Although networks can comprise a broad range of molecular types (e.g., proteins, chemicals, small molecules, etc.), they are generally reduced to the most direct outcome of our genetic makeup (i.e., the genetic and protein levels) such that we can garner mechanistic insights on how they operate. Thus, pathways are frequently viewed and simplified to “gene sets”, the collection of all genes/proteins that constitute a pathway, due to the major challenges of incorporating complex network topology and translating the variety of relationships they comprise into pathway analysis methods. Although pathway network representations offer a comprehensive picture of the interactions occurring in a given pathway, limitations still exist such as incorporating kinetic or time information in biological reactions. To address these shortcomings, various algorithms and techniques have been developed to model and simulate the dynamic changes of a pathway both qualitatively and over time [8–12].

While pathways have been introduced as powerful resources to store knowledge, their capabilities extend far beyond data warehousing. During the last decades, pathway networks have also been extensively used to complement and assist in the generation of new hypotheses and the interpretation of biomedical data. They have now become one of the cornerstones of data-driven analyses in systems biology. There are several reasons that explain the extensive use of pathway-driven analysis [14]. First, pathways are often associated with familiar biological and medical concepts (e.g., inflammation, cell death, etc.), thereby simplifying and facilitating the interpretation and comparison of results. Second, they support drug identification and development by elucidating their downstream mechanistic effects. Third, they reduce complexity in a field involving millions of molecular entities (e.g., genes, chemicals, (SNPs), protein variants, etc.). Thus, they indirectly act as a dimensionality reduction technique by projecting results onto a smaller shared feature space. Finally, due to their inherent multi-scale nature, they enable the integration of multiple *-omics* data (e.g., metabolomics, genomics, and proteomics). Taken together, the use of pathway constructs opens the door to not only better understanding of biology, but also to novel approaches aimed at drug

---

<sup>1</sup>Hereafter, the term “network” will be used interchangeably with the term “graph”.



**Figure 2: A)** MERTK signaling pathway. MERTK is a receptor tyrosine kinase that transduces signals into the cytoplasm after the binding of several ligands such as GAS6, Protein S, Tubby (TUB), TULP1 and LGALS3. The downstream effects of MERTK activation range from regulating processes such as cell survival or migration to cell differentiation and apoptosis. This figure was adapted from [13]. **B)** Pathway representation as a network. Biological entities are represented as nodes, and their interactions as edges.

identification, precision medicine, and disease modeling.

One of the current challenges in systems biology is in defining the boundaries of these modular units that we call pathways. It is difficult not only to identify the set of interactions comprising a pathway, but also to demarcate the limits of where a pathway starts and/or where it ends. Answering this question is not a trivial task



due to pathway crosstalk (i.e., pathways with up- or down-stream effects on each other, such as feedback loops) and the involvement of genes in multiple pathways (i.e., pleiotropic genes). Although this question leads to philosophical discussions around the nature of "what really is a pathway?", these questions are often ignored because pathways are inherently abstract concepts defined by researchers based on current scientific knowledge. Accordingly, pathway demarcations are dynamic and change over time in parallel with scientific developments. Further, investigating the boundaries of a particular pathway is a time- and labor-intensive task. First, a researcher must manually investigate the literature to formulate a pathway. Next, in order to prove her hypotheses, she must conduct dedicated experiments varying from classical knock-out to advanced gene editing techniques such as CRISPR/Cas9 that aim at elucidating the downstream effects of a pathway. In summary, characterizing new pathways and establishing their borders is a challenging task that requires significant amounts of resources. These resources must nonetheless be invested in order to gain a comprehensive overview of the pathway landscape.

## 1.2 Pathway databases

From the end of the last century, several efforts from various research groups, institutions, and private companies have focused on capturing disparate facets of pathway knowledge (e.g., signaling cascades, metabolic routes, and regulatory networks) and storing them in databases aimed at organizing information from this domain. According to PathGuide, there exist about a thousand pathway databases available<sup>2</sup> [15, 16]. One of the reasons explaining their rapid growth was through a need to formalize and store information generated by the explosive growth of the biomedical literature. However, this large number of databases also implies that these databases have been independently implemented and are currently isolated in so-called "data silos", thus hampering centralization approaches that seek to consolidate their knowledge.

While there exist hundreds of databases, only a handful of them are highly cited and employed (**Table 1**). There are several reasons that could explain this. First, the majority of databases are limited in scope with regard to the number of pathways they cover or they present outdated pathway representations since

---

<sup>2</sup>Note that this is an approximation intended to provide a rough estimate on the current number of databases. Moreover, it is important to mention that the last update in PathGuide was conducted in September of 2017, while this thesis was written in 2019.

| Type        | Pathway Resource | Publications |
|-------------|------------------|--------------|
| Primary     | KEGG             | 27.713       |
|             | Reactome         | 3.765        |
|             | WikiPathways     | 651          |
| Integrative | MSigDB           | 2.892        |
|             | Pathway Commons  | 1.640        |
|             | ConsensusPathDB  | 339          |

**Table 1:** Number of publications citing major pathway resources for pathway enrichment in PubMed Central (PMC), 2019. It is important to note the difference between primary (i.e., resources containing their own pathway information) and integrative databases (i.e., resources that integrate information from multiple databases). The latter are also referred as meta databases in literature. To develop an estimate on the number of publications using several pathway databases for pathway enrichment, SCAIView (<http://academia.scaiview.com/academia>; indexed on 01/03/2019) was used to conduct the following query using the PMC corpus: “<pathway resource>” AND “pathway enrichment”.

these resources can demand extensive manual curation. Second, the recognition of certain resources as reference databases as well as the preference of the researcher conducting the study introduces a bias towards the use of certain databases. Third, the funding body of each database (i.e., academic institutions vs. private companies) directly influences whether the access is public or not and thus, its usage. Below, a survey of the major resources in the field is presented.

- **Kyoto Encyclopedia of Genes and Genomes (KEGG).** As one of the oldest databases in the field, KEGG comprises a collection of pathway-related materials including networks, genomic information, and schematic representations for hundreds of pathways and metabolic routes in different species [17]. This resource has been maintained since 1995 by Kanehisa’s laboratory at Kyoto University in Japan. The main asset of KEGG is in its set of manually drawn pathway maps, representing molecular interaction and reaction networks. These are divided into several sections depending on their function or nature: metabolism, genetic information processing, environmental information processing, cellular processes, organismal systems, human diseases, and drug development [17].
- **Reactome.** This database is one of the largest public resources for biological pathways [18, 19]. Reactome is curated and maintained by an international multidisciplinary team by institutions from Canada, the United States, and

Europe since 2003. As its name suggests, Reactome's modeling unit is a biological reaction where each reactant and product is linked to its corresponding reaction. Thus, the aggregate of reactions constitute a network of biological interactions that are then grouped into pathways. The latest release of the database contains several thousand pathways for 79 species, including two thousand pathways for *Homo sapiens* alone<sup>3</sup>. Furthermore, its powerful tools, like its pathway browser, enable the scientific community to exploit the information in this resource by exploring pathway networks, overlaying data, and conducting pathway enrichment analysis, among other tasks. These tools are complemented by a dedicated Application Programming Interface (API) that offers downloading the database content in disparate formats as well as making complex queries to the database. Additionally, it is important to remark that the content of this database is not only highly curated but also cross-referenced and linked to other databases using controlled ontologies.

- **WikiPathways.** This resource is a community-driven database for contributing and maintaining content dedicated to biological pathways [20–22]. While the core of WikiPathways is comprised of peer-reviewed pathways, any registered user can curate and submit pathways to this resource, thus facilitating both outreach and its maintenance. Furthermore, it contains large amounts of Reactome content thanks to a recently implemented converter [23]. Although WikiPathways contains pathways for multiple species, its main asset is the collection of approximately 500 human pathways that have been made public to the community through its open access web application.
- **Gene Ontology (GO).** Despite the fact that it is not technically a pathway database, this resource provides a hierarchically organized set of thousands of standardized terms for biological processes, molecular functions and cellular components, as well as curated and predicted gene annotations based on these terms for multiple species. Its annotations can be used to interpret genomic information by asking questions such as how, where, and in which context a gene or protein operates. Additionally, GO is complemented by other databases such as PANTHER [24]. Therefore, GO is also commonly used for functional enrichment analysis. Although this resource does not yet contain pathway networks as the previously mentioned resources, it is important to remark that GO has proposed a new syntax for joining its annotations into larger models of biological function that could represent pathways, as will be discussed in the next section. In summary, GO can be

---

<sup>3</sup>Statistics based on Reactome's release number 68.

used for pathway enrichment analysis for a comprehensive representation of multi-scale relationships across biological entities.

While pathway databases cover a variety of scopes (e.g., metabolic or signaling) and contexts (e.g., cellular- and species-specific databases), the majority of studies thus far only employ a single database (**Table 1**). There could be two reasons that might explain this. First, researchers usually do not require specialized databases but rather generalized ones that cover as much pathway knowledge as possible. Second, running an analysis on a different database essentially means duplicating the workload, as analytic tools can be run with just one format. Further, this limited interoperability across tools and databases has been magnified by the adoption of multiple standards. Consequently, integrative efforts have continuously attempted to consolidate disparate databases, aiming to centralize pathway information.

Consolidating the knowledge contained in various databases is typically conducted by a so-called meta database (i.e., a database of databases). One of the most well-known meta databases is Molecular Signatures Database (MSigDB). This resource is a collection of publicly available gene sets annotated to their corresponding pathways. Other popular meta databases such as Pathway Commons [25] or ConsensusPathDB [26, 27] go one step further by accommodating pathway networks from multiple resources. Furthermore, to enable the exploration of pathway topology, they are complemented by corresponding web applications. Nevertheless, despite the use of these meta databases being especially suited for analyzing consolidated pathway information, their underlying merged data is not completely harmonized nor linked. For instance, because Pathway Commons does not harmonize the interactions from original resources, it is not possible to investigate the consensus or crosstalk of two overlaid networks from disparate resources. Additionally, because related pathways across resources have never been annotated and linked together, a typical pathway enrichment analysis could yield duplicate pathways (e.g., Pathway A from resource X and Pathway A' from resource Y).

### 1.3 Interoperability and integration of pathway databases

While semantic web technologies have paved the way for integrative approaches to manage, retrieve, represent, and harmonize knowledge, there exist two fundamental challenges that have impeded the path to make pathway knowledge fully interoperable across databases. The first barrier, as previously mentioned,

is related to the abstract nature of pathway delineations. This, together with the absence of a dedicated pathway controlled vocabulary until recently [28, 29], explains why there are no pathway cross-references and mappings across databases. Similar to the lack of controlled vocabularies during the first decades of database development, the absence of a golden standard to formalize pathway knowledge led to the advent of multiple formats and schemata. However, all these novel formats share a fundamental principle: they are all computable formats which prioritize human readability in order to facilitate the work of curators. While the existence of heterogeneous standards offer researchers numerous alternatives to implement databases depending on their purpose or the underlying data to be stored, they also pose a technical obstacle when harmonizing data across distinct resources. The following section presents a survey of standard formats used to formalize pathway data.



**Figure 3:** Diversity in formats used by the four pathway databases reviewed in this thesis. Although the majority of these resources export to more than one standard, a limited number of them are shared across resources.

- **Resource Description Framework (RDF).** This format is a standard format for storing, managing, and modeling knowledge and it originated in the semantic web domain. It was designed to describe resources and the relationships that link them. RDF is comprised of triples, each formed by a subject, a predicate, and an object, in which the subject is the acting resource, the predicate is a linking relationship, and the object is the resource that is acted upon. Both subject and object can be represented as a Uniform Resource Identifier (URI), the base of its vocabulary. This flexibility permits the merging of data,

though the schemas which form their basis may differ in contrast to other formats such as Extensible Markup Language (XML). Further, the use of triples as semantic units supports linking data across distinct resources, as illustrated by Bio2RDF [30], WikiPathways [31], and Scholia [32].

- **Biological Pathway Exchange (BioPAX).** This format was initially designed to drive the exchange of biological pathway data, thereby facilitating its integration, visualization, and analysis [33]. Further, it is highly effective in handling ontologies and exporting its content to other data types since it is derived from two semantic web standards, RDF and Web Ontology Language (OWL). BioPAX 3.0, its latest version, defines five top level classes (i.e., entities, genes, physical entities, interactions and pathways) to support the representation of pathways. The ontology defines discrete physical entities, interactions as sets of physical entities and pathways as sets of interactions. In a graph representation, this would be analogous to nodes, hyperedges and graphs, respectively. Numerous databases use BioPAX to store pathway knowledge, including Reactome, WikiPathways as well as Pathway Commons (**Figure 3**).
- **Systems Biology Markup Language (SBML).** This format, which is based on XML, was designed to represent computational models of systems biology [34]. Although SBML was originally designed to serve as a lingua franca in the field of biochemical network modeling, it has evolved to represent other biological processes. Due to its origins, this language offers users the option to include quantitative information in the form of equations such as chemical reactions. This promotes the exchange of quantitative models of biochemical networks between different simulation tools. Physical entities are denoted species and processes are called reactions. They can be encoded as models, that when decomposed, closely resemble chemical reaction equations. Finally, SBML is used or can be exported by various databases such as Reactome (**Figure 3**) and HumanCyc [35].
- **Biological Expression Language (BEL).** Conceived in the private sector, this language is specially suited to represent biomedical knowledge in a computable form by capturing causal and correlative relationships [36]. BEL allows for the inclusion of a minimal set of information for each triple or BEL statement (i.e., a reference, evidence text, and defining entities according to the functions or relationships allowed in the language). This set of triples of the form subject, predicate, and object are then combined into a network. Furthermore, entities are formalized by using external vocabularies and ontologies, thus easing their normalization and cross-reference to domain-specific databases (e.g., Chemical Entities of Biological Interest

(CHEBI) [37], HUGO Gene Nomenclature Committee (HGNC) [38], etc.). Additionally, its inherent flexibility supports annotating triples with contextual information as well as encoding entities spanning multiple scales (e.g., molecules, cellular processes, phenotypes, etc.). BEL is now open source and it is being developed by a consortium of institutions [39] that provide tools and resources to visualize and analyze the resulting networks (e.g., BEL Editor, Knowledge Assembly Model (KAM) navigator, PyBEL [40]). Similar to Bio2RDF, the Bio2BEL framework [41] demonstrates how BEL can drive semantic integration and harmonization in networks and systems biology.

- **Other standards.** Although the aforementioned standards are backed by larger communities, other formats extensively used in the field also exist. For instance, Systems Biology Graphical Notation (SBGN), is suited for the storage and exchange of signalling pathway, metabolic network and gene regulatory network information [42]. Further, Proteomics Standards Initiative - Molecular Interaction (PSI-MI) is a data exchange format for molecular interactions maintained by the Human Proteome Organization (HUPO) [43], and Simple Interaction Format (SIF) is an elegant format designed to build graphs from lists of molecular interaction units. Two other XML-based languages, CellML [44] and KGML, are respectively designed to describe mathematical models and pathway maps in KEGG. Lastly, GO has recently developed a new format called Causal Activity Models (CAM) designed to give more expressibility to its annotations and convert them to networks (**Figure 3**).

While most of the standard languages described in this survey share capabilities and have been proven to effectively model biological knowledge, each language is best suited for a particular application depending on both the goal and domain of study. For example, SBML specializes in modeling quantitative aspects of molecular processes, including chemical kinetics. On the other hand, both BEL and BioPAX have a strong focus on capturing interactions across biological entities. However, their structural differences influence how flexible curators can be in representing biological entities and their interconnections. Since the structure of BEL more closely resembles a generic network, it allows for more freedom in defining relationships and entities. This enables assembling contextualized knowledge from multiple scales (e.g., molecular, phenotypic, and genetic level), thus, making it particularly well-suited for clinical applications and disease modeling. However, this may cause harmonization issues if two curators represent entities differently. BioPAX, on the other hand, has a more complex structure that encourages curators to define entities using standard biological paradigms that can make it highly

verbose. Moreover, some formats offer curators a predefined vocabulary to express relationships (e.g., BEL) while others such as BioPAX let curators decide their own. In terms of usage by the bioinformatics community, BioPAX and SBML are supported by a larger number of software tools and databases than BEL and RDF. Ultimately, all languages mentioned in this thesis have been designed to connect entities or relationships to external vocabularies in order to facilitate the cross-linking and transforming of knowledge from one language to another. This eases the burden of pathway knowledge exchange by integrating resources that use different formats, thus, connecting data silos.

The properties and characteristics of a particular database format, alongside the complementary software tools that support it, play an important role for the adoption and application of a given pathway resource. We can divide these tools into three different categories depending on their purpose: (i) curation workflows, (ii) analytical tools, and (iii) parsers and converters. Among the noteworthy in the first category are Payao for SBML [45], MINERVA or the SBGN editor for SBGN [46, 47] and NaviCell for any XML-based format [48]. Thanks to the compatibility across formats, the second category is broader and offers numerous visualization and distribution tools, such as NDEx or PathVisio [49–56]. Finally, the latter group accounts for tools designed to convert from one format to the other such as [23, 51, 57–60]. These tools are ultimately responsible for enabling interoperability across resources. Converters operate by applying a set of inference rules to map two distinct data models, which effectively transforms one format to another. However, conducting this mapping task often leads to the inclusion of ambiguities, redundancies, or even information loss. Summarizing, this harmonization challenge necessitates converting each of the database formats into a consensus schema that integrate their heterogeneous information.

The wide range of both databases and formats complicates evaluating the potential overlap across pathway databases. Furthermore, the presence of database-specific terminologies and formats compels manual intervention in order to assess the consensus of a particular across databases as outlined by [61]. To integrate multiple databases into centralized repositories, different approaches have attempted to consolidate disparate databases by converting each of their individual formats to a common structure.

Pathway Commons, one of the meta databases previously mentioned, has undertaken this tremendous effort of uniting databases together with the help of BioPAX [25]. OmniPath, on the other hand, combines multimodal information from heterogeneous databases (e.g., transcription factors, protein-protein interactions, etc.) and assembles it to a simplified triple-based format [62]. Furthermore,



this resource is complemented with a Python package that facilitates its usage for other applications. Similar to OmniPath, the graphite R package integrates multiple sources and enables users to manipulate the resulting networks [63]. However, as previously mentioned, there are some limitations of these integrative approaches. First, the fact that OmniPath and graphite do not follow a systems biology standard, but rather implement generic network schemata, leads to an over-simplification of relationships present in the resources they integrate. For example, OmniPath does not include directionality, though this information is present in most of its original resources. Furthermore, their underlying networks exclusively contain signed information (i.e., activation and inhibition) and lose contextual information such as differentiating between biological classes (e.g., gene versus protein) or how activation is mediated (e.g., phosphorylation, biochemical reaction, etc.).

While integrative resources facilitate the generation of multi-scale knowledge graphs, data integration has to be conducted with a minimal loss of information. Capturing contextual information is essential for analyzing *-omics* data with the support of the knowledge embedded in the network structure. This, together with the rapid development of novel machine and deep learning techniques [64–66] calls for sophisticated approaches that adequately harmonize both biological entities and relationships, while permitting the contextualization of the information comprised in the knowledge graph. The next chapter introduces the concept of disease maps (i.e., a knowledge graph of a given disease) and how they can be employed to represent the mechanisms around human disorders.

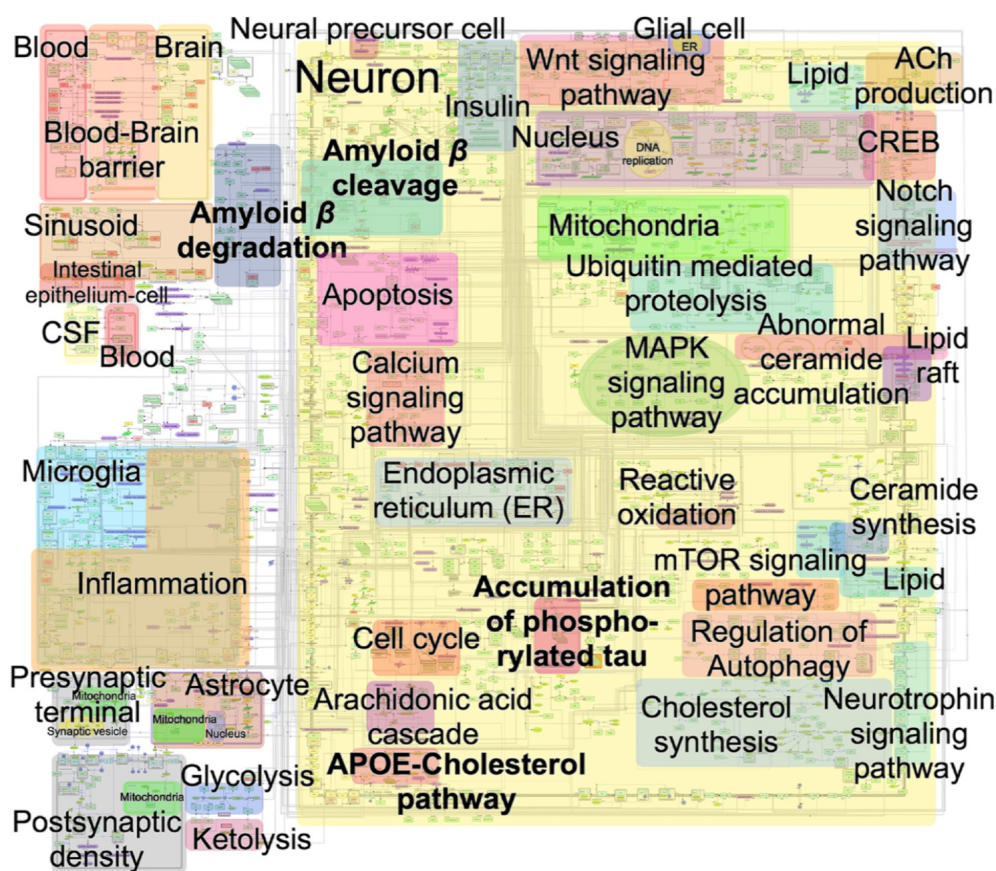
## 1.4 Disease maps

Canonical pathways formalize sets of "biological snapshots" that correspond to the chains of causation occurring in normal cellular physiology. However, pathway interactions can be altered by their environment and context [5]. In other words, the same pathway in the same organism can effectively yield two completely opposite outcomes in two different cellular types (e.g., neuron, adipocyte, etc.) or states (e.g., age, signaling from neighbouring cells, nutrition, cell cycle stage, etc.). Therefore, establishing clear and delineated pathway boundaries based on contextualized information is crucial to better comprehend and interpret the inherently dynamic nature of biology. Accordingly, classical pathway-centric approaches must be extended to incorporate contextual information in order to shape and adapt pathway knowledge depending on the context given in the studied model.

Contextualization might explain the success of pathway resources to decipher and unravel the underlying pathophysiological mechanisms in certain diseases, especially in those where research has been abundant (e.g., cancer and metabolic disorders), and the lack of success in others (e.g., neurodegenerative or psychiatric conditions). The latter diseases tend to be especially complicated due to their still unknown multifactorial nature. This, in turn, has translated into a limited number of treatments for them (if any).

Capturing disease-specific information is essential because pathways can have different behaviours depending on one or the other scenario. For this reason, roadmaps were launched to build disease maps for various conditions organized by the Disease Maps Project [67]. The goal of disease maps is to formalize the knowledge around signaling, metabolic, and gene regulatory pathway networks that are involved in the disease of study in order to reveal underlying crosstalk and interplays across disease mechanisms. This task requires both clinicians and biologists to curate relevant literature in order to ensure that key molecular players involved in the disease pathophysiology are present. Moreover, as novel hypotheses or mechanisms are proposed, the content has to be adequately updated, ensuring that the new pieces are coherently integrated in the "disease map puzzle". As opposed to standard pathway resources, disease maps not only contextualize disease-specific information but often add several other biological aspects and scales such as Single Nucleotide Polymorphism (SNP)s, gene variants, and clinical phenotypes associated with the condition [68, 69]. Therefore, disease maps go beyond classical pathway representations by integrating novel biological scales and mechanistic information to provide a more comprehensive overview of the disease landscape. In summary, disease maps are computable assemblies of expert-curated and contextualized knowledge that can not only be used to store this information but also to model disorders and generate new hypotheses.

Over the last few years, several initiatives were launched to build disease maps in conditions such as Alzheimer's disease (AD) (AlzPathway), Parkinson's disease (PD) (PDMap), asthma (AsthmaMap), cancer (Atlas of Cancer Signalling Network), rheumatoid arthritis, and influenza [70–76]. Apart from the continuous updates of these existing resources, other disease maps are also currently under development in areas such as acute kidney injury, spinal cord injury, Meniere's disease, lung cancer, and cystic fibrosis, among others [77]. Of the above mentioned disease maps, among the largest are AlzPathway (**Figure 4**) and PDMap, developed by two particular efforts in the field of neurology. Although mechanistic information is lacking in this challenging area, both resources emerged as comprehensive catalogs of their respective conditions by incorporating information from over a hundred review articles in the case of AlzPathway, and over thousand research articles



**Figure 4:** Overview of AlzPathway overlaid with canonical pathway annotations. Most of the pathways identified in this work overlap with NeuroMMSig. This figure was taken from [70].

in the case of PMap. Further, PMap is complemented by MINERVA, a web application that supports the curation, annotation and visualization of biological networks [46]. On the other hand, AlzPathway cannot be directly explored on its website but rather must be visualized with the help of auxiliary software (**Figure 4**). Although both maps can be explored through user-friendly interfaces that even show cell compartmentalization, the format chosen (i.e, SBML), constrains scientists to analyze and investigate network crosstalk since an entity can be present multiple times in the map. In other words, because the map layout is compartmentalized and enables presenting multiple representations of the same molecule in different mechanistic or pathway networks, these networks cannot be later overlaid without processing the original networks. This duplication issue at the node level can only be overcome with substantial manual effort (i.e., manually

linking entities in a post-processing step) or by converting such formats to other graph-generic formats.

Concluding, contextualizing and formalizing knowledge in the form of disease maps enables cataloging crosstalk across molecular players and pathways in a particular disease. By doing so, the analysis of disease modeling supports investigating disease aetiology by generating novel mechanistic hypotheses. However, the expansion and maintenance of disease maps is crucial in order to continue integrating the knowledge coming from novel literature. Furthermore, in the particular case of neurological disorders, disease maps could incorporate other aspects and biological scales related to the condition such as imaging readouts (e.g., volume of region brains), biomarkers, and clinical features (e.g., psychological tests). Finally, future overarching approaches that connect these computable knowledge templates to real multi-scale and multimodal cohorts could shed some light on the mechanisms underlying aetiology of these complex disorders.

## 1.5 Neurodegenerative and psychiatric disorders

Neurological disorders group together a series of conditions, such as Alzheimer's disease, Parkinson's disease, epilepsy, or multiple sclerosis, where there exist nervous system malfunctions or damage. On the other hand, psychiatric disorders, such as anxiety, schizophrenia, or post-traumatic stress disorder manifest through disturbed behaviour and emotional states. These two groups of disorders impose a major economic and social burden. Not only the patient, but also the family and caretakers of the patient are profoundly impacted by the decline of the patient, and the accompanying emotional burden. On the other hand, in economic terms, dementia alone has a global impact larger than one trillion dollars in the United States alone [78]. Furthermore, the population growth expectations suggest that the economic costs associated with mental illnesses will grow exponentially over the next 30 years [79]. The following subsections introduce the three neurodegenerative and psychiatric conditions (i.e., PD, AD, Post-traumatic stress disorder (PTSD)) that this thesis focuses on.

### 1.5.1 Alzheimer's disease

Alzheimer's disease (AD) is a neurodegenerative disease that progressively affects memory, thinking, and behavior by inducing neuronal dysfunction. This condition is the most predominant form of dementia and is the neurological disorder with the highest prevalence in the population [80, 81]. Although multiple hypotheses have been proposed [82–86], little is known about its multifactorial aetiology. The variety of mechanisms implicated (or thought to be) and the vast number of possible chemicals to target them can explain why, despite the billions invested by pharmaceutical companies, there still is no cure for AD, only treatments that relieve patients from their symptoms [87]. Other reasons could be attributed to the fact that patients are treated in advanced stages of the disease (i.e., treatment comes too late) [88] or trials are conducted in highly heterogeneous patient groups (i.e., drugs might work exclusively in a subpopulation).

Today, it is estimated that about 50 million people live with some form of dementia. The majority of the cases exist in developed countries where patients can be diagnosed and have access to health care. By 2050, when the population pyramids of developing countries evolve from their current expansive shapes (i.e., bell-curved) to stationary ones (i.e., rectangular shape), this number is expected to be tripled [89]. However, the number of scientific publications related to dementia is ten times smaller than the cancer field [89]. The combination of this underrepresentation together with the future demographic outlooks could explain why western countries have set dementia as a public health priority and have launched numerous projects addressing this issue.

Although there exist multiple types of dementia, there tends to be agreement in the literature on the subdivision of AD into two main subtypes [90–92]:

- **Familial AD.** This subtype is related to mutations involved in AD-related genes, such as APP, PSEN1 and PSEN2. New insights on these etiological agents are essential for a better understanding of the pathogenesis of AD.
- **Sporadic AD.** Accounting for about 95% of all cases, sporadic AD presents the same symptoms as the previous subtype though it cannot be distinguished from the familial form since the etiology of this form has yet to be fully elucidated. It is believed that it is caused by environmental factors as well as a genetic component.

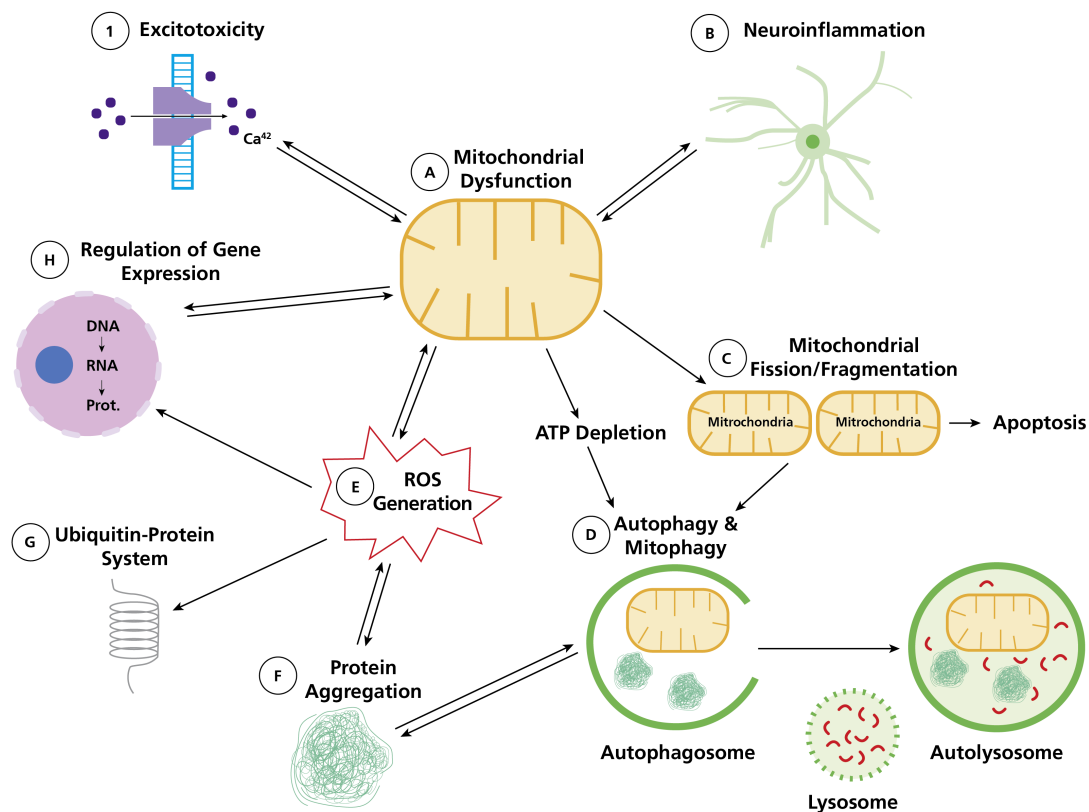
## 1.5.2 Parkinson's disease

Parkinson's disease (PD), the second most common neurodegenerative disease, is characterized by a series of unknown detrimental changes in the central nervous system that lead to dysfunction in the motor system. PD pathophysiology is associated with a deterioration of the dopamine release system that ultimately disrupts motor system skills, translating into unstable and unplanned movements. Hence, trembling movements are the most common symptom at early stages. Later stages, however, develop into cognitive decline and behavioral issues as the areas of the brain become affected [93]. Unfortunately, although some of the symptoms can be alleviated, as of yet there is no cure for PD.

Epidemiologically, PD is a highly prevalent condition as studies indicate approximately ten million patients are affected worldwide [94]. However, this number is expected to double in the next decades due to an increase in aging populations, longer disease durations, and environmental as well as social risk factors [95]. Conversely to AD, this condition is more prevalent in men with a 1.6 to 1 ratio [96]. This difference is believed to be attributed to the neuroprotective effect of estrogen in women [97–99].

Since PD is a multifactorial condition and its pathophysiology has yet to be fully understood, various studies have been conducted and found numerous mechanisms to be associated with PD (**Figure 5**). According to the possible etiology and clinical implications, two subtypes of PD can be characterized [102]:

- **Familial or monogenic PD.** Accounting for approximately 5% of the diagnosed cases of PD, this subtype is caused by inheritable monogenic genetic variants, such as SNCA, PINK1, and LRRK2 [103]. Typical traits of this PD subtype are both early onset (around forty years) and accelerated disease progression.
- **Idiopathic or sporadic PD.** Idiopathic PD constitutes roughly 90% of diagnosed PD cases. Men of age 80 years represent the majority of the cases and the average age of diagnosis is 55 years old. In contrast to familial PD, the pathogenesis of this type is gradual and its pathophysiology is associated with epigenetic and environmental factors [104].



**Figure 5:** Schematic representation of the crosstalk between different mechanisms implicated in PD pathophysiology. Both Mitochondrial dysfunction (A) and neuroinflammation (B) result in a cascade of cellular events that lead to apoptosis such as generation of Reactive Oxygen Species (ROS) (E), mitochondrial fission/fragmentation (C) or ATP depletion. Cellular responses to these changes include alteration in gene expression (H) or autophagy and mitophagy (D). These processes are related with the aggregation of proteins such as synuclein and activation of the ubiquitin system, both of which are disease hallmarks. Finally, excitotoxicity (I) caused by a dysregulation in the influx of Ca<sup>2+</sup> is also related to mitochondrial dysfunction through the depolarization of its membrane. This figure has been adapted from [100, 101].

### 1.5.3 Post-traumatic stress disorder

PTSD is a common psychiatric disorder that can occur in individuals after a traumatic event [105]. This condition is diagnosed by psychologists based on the presentation of four characteristic symptom: intrusions, avoidance, negative

cognitions/mood, and hyperarousal [106]. While PTSD pathophysiology is not yet fully understood, research suggests that numerous neurological systems that regulate mental and physical health functions are implicated [107]. Furthermore, PTSD symptoms complicate its diagnosis. In fact, it was not officially recognized as a condition until 1980 by the American Psychiatric Association [108].

From the epidemiological point of view, [105] has a prevalence around 3.5% in the United States [106]. Globally, although trauma exposure is unequally distributed due to cultural differences and the presence of local conflicts, this condition is present in around 10% of the total population at some point of their lives [109]. Finally, it is important to note these figures can be considered conservative as epidemiological studies indicate that over 70% of the population experiences a traumatic event [109, 110].

Besides the epidemiological figures, PTSD has a significant impact on our economy and society. Economically, the costs derived from this condition are related to the health resources used, such as medication, and resources lost in terms of productivity and presenteeism. Due to the difficulty in estimating economic cost-of-illness, there have been no studies focusing on the global economic impact of PTSD. However, a local study based in Northern Ireland reported surprisingly large figures for such a small region [111]. Similarly to neurodegenerative diseases, the social impact is not restricted to patients but also to families, relatives, and the care staff who suffer from prolonged stress, depression and other psychological disorders.

The principal treatments for PTSD patients are psychotherapy and medication. The most common prescriptions are antidepressants that act as a Selective Serotonin Re-uptake Inhibitors (SSRIs) (e.g., Zoloft and Paxil). However, the mechanism of action of these drugs is still unknown, and such generic medication is prescribed for multiple other psychiatric disorders. Furthermore, the benefits of these drugs may be outweighed by their numerous side effects [112]. Since we still lack the mechanistic understanding about the pathophysiological changes occurring in the brain that lead to this disorder, it is critical to start by analyzing biomarker data in order to pinpoint endophenotypic traits implicated in this disorder. In addition, biomarker discovery, as in neurodegenerative diseases, is essential for detecting symptomatic patients at an early stage of the disease so they can be immediately treated at a early stage of the disease for more timely treatments.



## 1.6 Translational research: applying knowledge-derived hypotheses to the clinic

The previous section illustrates the necessity to elucidate the pathophysiology underlying these disorders. While knowledge-driven approaches can be used to exploit pathway and mechanistic information and model the disease, such tasks have to be complemented with data-driven approaches. One of the classical examples on how the crosstalk between the two is essential for driving science is the drug development process. In this domain, data-driven approaches are applied to validate a candidate drug by analyzing data from a clinical- or cohort-based study. However, to be successful, a study needs to be designed in a way that takes into account prior knowledge (e.g., reflecting patient heterogeneity and conducting a meta-analysis of the literature).

Disease have a time dimension that also needs to be modelled by knowledge-driven approaches. This aspect is well-characterized in longitudinal studies whose data can be analyzed to study patient-specific progression. Using this information, we can stratify the patients that present similar patterns during disease progression and analyze the pathways or mechanisms that differentiate these patient subgroups. This, in turn, can support us understanding how their mechanistic characteristics lead to disparate clinical phenotypes. Building such a "mechanism-based taxonomy" is crucial to reveal the mechanistic underpinnings of highly heterogeneous patient populations in conditions whose pathophysiology is yet unknown.

Stratification approaches are especially relevant for the idiopathic subtypes of AD and PD since the majority of the disease population falls into heterogeneous groups, as previously discussed. Specific patient subtypes must be properly characterized in order to correctly identify the disease mechanisms at work. Without this crucial first step, clinical trials will fail. This illustrates the power of a combined data- and knowledge-driven approach for knowledge discovery and its broad applicability to precision medicine and translational research.

Crossing the translational divide between knowledge-driven discovery and clinical implementation first requires linking information from the biomarkers and endpoints measured in a clinical study with knowledge-derived models. This crucial step involves curating and organizing this information in order to facilitate the harmonization and integration of results from multiple studies. Therefore, numerous initiatives currently aim to catalogue biomarker information on particular

conditions, such as colorectal cancer [113], Alzheimer's disease [114], tuberculosis [115], and liver cancer [116].

While the value of these integrative efforts is often underestimated and is associated with demanding tasks such as data preprocessing and harmonization, these resources foster research by providing a more comprehensive view of the information available. For instance, making large studies such as Alzheimer's Disease Neuroimaging Initiative (ADNI) and Parkinson's Progression Markers Initiative (PPMI) interoperable allow for replicating and validating previous studies. In addition, the considerable amount of data generated by merging studies that share significant overlap enables developing more robust models and drawing and validating new conclusions and hypotheses.

## 1.7 Outline of the thesis

This thesis first focuses on the development of novel software tools and web applications designed to better interlink, consolidate, and harmonize knowledge across different pathway databases. Chapter 2 presents ComPath, an ecosystem that supports curation of pathway mappings between databases and fosters the exploration of pathway information through several novel visualizations. By using this ecosystem, we curated a novel dataset of pathway mappings that provides a comprehensive view on pathway relationships across three major databases (i.e., KEGG, Reactome, and WikiPathways). Chapter 3 presents PathMe, the first framework which successfully harmonizes pathway networks across the previously mentioned databases, at both entity and relationship level. Both tools are complemented with their corresponding web applications facilitating the exploration and analysis of the knowledge they consolidate. Finally, chapter 4 presents a comprehensive benchmarking of individual pathway databases on statistical enrichment analysis and predictive modeling methods. Furthermore, with the help of the former two tools (i.e., ComPath and PathMe), we establish an approach to integrate pathway knowledge from different resources into a merged dataset to demonstrate that integrative approaches outperform individual databases. This study illustrates how database choice has a significant impact on results and highlights the importance of integrative approaches as a way to mitigate this bias.

The following chapters outline knowledge- and data-driven approaches aiming to unravel the underlying pathophysiological mechanisms involved in psychiatric

and neurodegenerative disorders. Chapter 5 presents an innovative enrichment paradigm, NeuroMMSig, supported by over 200 disease-specific mechanistic networks for three neurodegenerative disorders, as opposed to canonical pathways, to offer the scientific community a novel resource for knowledge discovery in the context of three conditions (AD, PD, and epilepsy). Chapter 6 introduces the first biomarker database in the context of PTSD. This resource, the first of its kind, catalogs biomarker information in a comprehensive database complemented by a web application aiming to facilitate future analysis and research in the field. Finally, chapter 7 illustrates how the crosstalk between machine learning predictive models derived from the major AD clinical study like ADNI, and knowledge-driven approaches such as NeuroMMSig can reveal promising mechanistic links in this condition.

The final chapter outlines the main topics, discusses the limitations and presents possible future directions of this work, serving as a conclusion of the thesis.



# 2 ComPath: an ecosystem for exploring, analyzing, and curating mappings across pathway databases

## Introduction

The growth of pathway knowledge that has accompanied the recent explosion of high-throughput biological data has led to the development of dozens of databases. However, the lack of interoperability between them hampers an integrative approach that can synergistically exploit these resources in a coordinated fashion. Due to the lack of a gold standard in the field of systems biology to represent pathways, various formats were adopted to improve reproducibility and facilitate the exchange of pathway knowledge [117]. Though several efforts have successfully accommodated multiple pathway databases, the absence of a unified pathway ontology [29] and the lack of inter-database mappings which impede the ability to assess the knowledge gaps and biases that may be present in pathway databases. This chapter presents a flexible software that is able to integrate gene-centric and chemical pathway data from multiple databases in order to explore, analyze, and curate pathway knowledge. Using this software, we established the first mappings across three of the major pathway databases (i.e., KEGG, Reactome, WikiPathways) [17, 20, 118].

Reprinted with permission from "Daniel Domingo-Fernández *et al.*. ComPath: an ecosystem for exploring, analyzing, and curating mappings across pathway databases. *npj Systems Biology and Applications*, Volume 4, 43, 13 December 2018". Copyright © Daniel Domingo-Fernández 2018.

## TECHNOLOGY FEATURE OPEN

## ComPath: an ecosystem for exploring, analyzing, and curating mappings across pathway databases

Daniel Domingo-Fernández<sup>1,2</sup>, Charles Tapley Hoyt<sup>1,2</sup>, Carlos Bobis-Álvarez<sup>3</sup>, Josep Marín-Llaó<sup>1,4</sup> and Martin Hofmann-Apitius<sup>1,2</sup>

Although pathways are widely used for the analysis and representation of biological systems, their lack of clear boundaries, their dispersion across numerous databases, and the lack of interoperability impedes the evaluation of the coverage, agreements, and discrepancies between them. Here, we present ComPath, an ecosystem that supports curation of pathway mappings between databases and fosters the exploration of pathway knowledge through several novel visualizations. We have curated mappings between three of the major pathway databases and present a case study focusing on Parkinson's disease that illustrates how ComPath can generate new biological insights by identifying pathway modules, clusters, and cross-talks with these mappings. The ComPath source code and resources are available at <https://github.com/ComPath> and the web application can be accessed at <https://compath.scai.fraunhofer.de/>.

*npj Systems Biology and Applications* (2018)4:43; <https://doi.org/10.1038/s41540-018-0078-8>

## INTRODUCTION

The notion of pathways enables the representation, formalization, and interpretation of biological events or series of interactions. Cataloging biological knowledge into pathways reduces complexity from all possible interacting molecular entities to a set of well-studied and validated functional relationships between molecular entities culminating in biological processes. Several efforts have generated databases of pathways with varying specificity and granularity that comprise signaling cascades, metabolic routes, and regulatory networks from precise signatures with no more than a couple of acting players to general pathways involving thousands of molecular players.<sup>1–4</sup>

Simplifying biology into pathways and representation as network models or mathematical models inevitably results in a loss of information such as spatiotemporal information or even entire biological entity types. The network abstraction facilitates pathway visualization and interpretation thanks to the harmony between biological networks and systems: nodes correspond to molecular entities and edges to types of interactions occurring between them (e.g., inhibition, phosphorylation, etc.). Although networks can comprise a broad range of molecular types (e.g., proteins, chemicals, small molecules, etc.), they are generally reduced to the most direct outcome of our genetic makeup - the genetic and protein levels - so that we can mechanistically understand their functionality. Thus, they are frequently viewed and simplified to “gene sets”, the collection of all genes/proteins that constitute the pathway, due to the major challenges of incorporating network topology and translating the variety of relationships into pathway analysis methods.

While dedicated research groups and commercial entities with experienced curators have led a majority of the efforts to compile, delineate, and store biological knowledge into pathway databases,<sup>2,5</sup> community and crowdsourced efforts have recently

gained traction.<sup>3,6</sup> Further, the variability in curation team composition, database scope (e.g., signaling pathways, gene regulatory networks, and metabolic processes), and curation guidelines led to the adoption of different (and in many ways incompatible) schemata and formalisms such as Biological Pathway Exchange (BioPAX;<sup>7</sup>) and Systems Biology Markup Language (SBML;<sup>8</sup>). These incompatibilities motivated the integration and harmonization of resources into pathway meta-databases such as Pathway Commons<sup>9</sup> and PathCards,<sup>10</sup> which focus on integrating databases; iPath,<sup>11</sup> which focuses on pathway visualization; and SIGNOR, which focuses on signaling pathways.<sup>12</sup>

Even after integrating multiple pathway databases into a pathway meta-database, it is difficult to assess the agreements, discrepancies, redundancy, and the complementarity of their contents because of the lack of availability of pathway mappings (e.g., pathway A from resource X is equivalent to pathway B from resource Y) in the original databases. These mappings are difficult to establish because of the arbitrary and overlapping nature of pathway boundaries as well as the absence of a common pathway nomenclature. Several controlled vocabularies have been generated as initial attempts to standardize pathway nomenclature,<sup>13,14</sup> but most pathway databases had already been established by the time these ontologies were published. Therefore, consolidating pathway knowledge is a persisting issue and it is still required to map pathways from different resources together to improve database interoperability.

Hierarchical clustering approaches have been presented as a way of grouping similar pathways based on their corresponding gene sets in order to propose pathway mappings.<sup>10,15</sup> Though these approaches can systematically cluster pathways from multiple resources, there are some limitations to consider: first, the usual tradeoff between over/under-clustering,<sup>16</sup> and second, pathway nomenclature and biological context are not considered

<sup>1</sup>Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing, 53754 Sankt Augustin, Germany; <sup>2</sup>Bonn-Aachen International Center for IT, Rheinische Friedrich-Wilhelms-Universität Bonn, 53115 Bonn, Germany; <sup>3</sup>Faculty of Medicine and Health Sciences, University of Oviedo, 33006 Oviedo, Spain and <sup>4</sup>Rovira i Virgili University, 43003 Tarragona, Spain

Correspondence: Daniel Domingo-Fernández ([daniel.domingo.fernandez@scai.fraunhofer.de](mailto:daniel.domingo.fernandez@scai.fraunhofer.de))

Received: 5 July 2018 Revised: 31 October 2018 Accepted: 2 November 2018

Published online: 13 December 2018

by the clustering algorithm; it often leaves out equivalent pathways with low similarity and ignores the context of the pathway (e.g., cell/disease specificity). Nevertheless, these limitations can be overcome by following clustering and prioritization methods with the manual curation required to interpret the abstract concepts that inherent to pathway definitions (e.g., biological process, cellular location, condition, etc.).

Though numerous algorithms<sup>17</sup> and tools<sup>4,18</sup> have been successfully applied to interpret experimental data through the context of pathway databases,<sup>19,20</sup> there has not yet been a systematic comparison between the contents of various pathway databases, an assessment of their overlaps and gaps, or an establishment of mappings. Previous studies have only focused on comparing a single or small set of well-established pathways across multiple resources.<sup>21,22</sup> For example, a comparison focused on metabolic pathways revealed how a set of five databases only agreed in a minimum core of the biochemistry knowledge.<sup>23</sup>

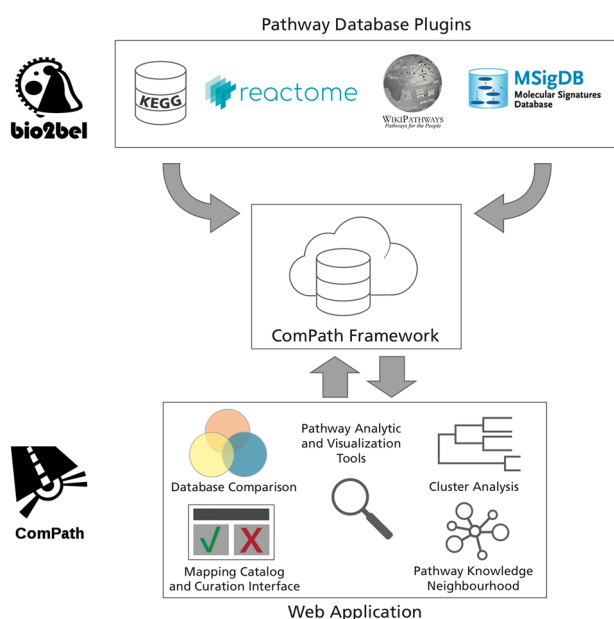
These studies demonstrate the need to connect insights provided by each pathway database to foster a greater understanding of the underlying biology. Here, we present ComPath, a web application that integrates content from publicly accessible pathway databases, generates comparisons, enables exploration, and facilitates curation of inter-database mappings.

## RESULTS

We developed an interactive web application that enables users to explore, analyze, and curate pathway knowledge. Below, we present three case studies illustrating how it can be used for each of these purposes. The figures for each were generated by interactive, dynamic views in the ComPath web application based on three major pathway databases: KEGG, Reactome, and WikiPathways (Fig. 1).

### Case study I: comparison of pathway databases

**Assessment of gene coverage.** Analysis of the overlaps between Kyoto Encyclopedia of Genes and Genomes (KEGG), Reactome, and WikiPathways revealed that there are ~3800 common human



**Fig. 1** The ComPath ecosystem has three main components: the pathway database plugins, the ComPath framework, and the ComPath web application. The ComPath framework mediates the communication between the plugins containing the pathway database information and the web application

genes shared between the three databases (Fig. 2a). While at least one common human gene was present in almost every pathway across each database, the number of pathways with more common human genes diminishes much more quickly in WikiPathways and Reactome (Supplementary Figure S1). This may be due to database properties such as pathway size (e.g., on average, pathways contain 90 genes in KEGG, 50 in Reactome, and 42 in WikiPathways) or gene promiscuity (i.e. genes functionally linked to many pathways) that might influence the results of analyses using pathway resources (Supplementary Table 2). For further investigation, the ComPath web application generates summary tables and creates several visualizations to enable exploration of the distributions of pathway size and gene memberships for each database, visualizations that present an overview of the database properties to help identify effects such as gene promiscuity or differences the distribution of gene set sizes (Fig. 2b).

**Exploration of pathways.** While the previous views produced gene-centric summaries of the contents of pathway databases, ComPath also enables the exploration of pathway similarity landscape using Clustergrammer.js.<sup>24</sup> Figure 2C illustrates how this view can identify clusters of pathways based on their similarity and then elucidate the hierarchical relationships between the Metabolic pathway, the largest KEGG pathway, and other more high-granular KEGG metabolic pathways (e.g., alpha-Linolenic acid metabolism, Lipoic acid metabolism, and ether lipid metabolism).

### Case study II: identification of pathway modules, overlaps, and interplays using pathway enrichment

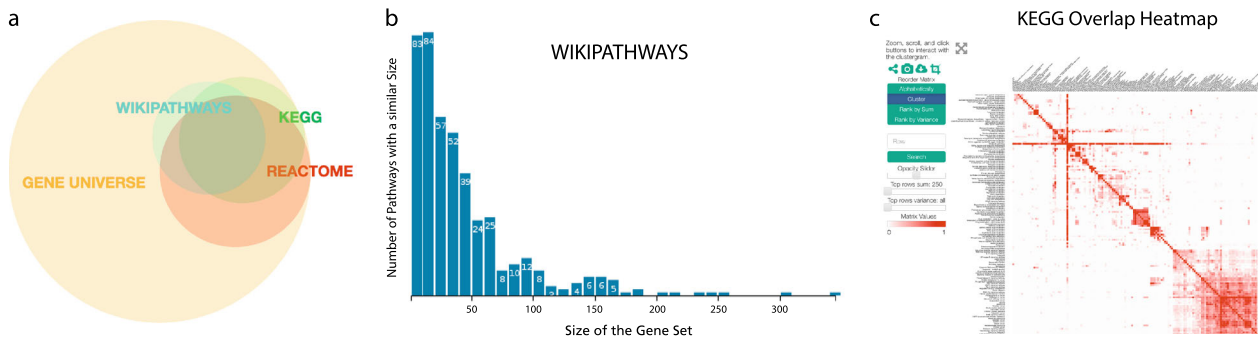
ComPath couples classic pathway enrichment analysis<sup>18,25–27</sup> with pathway-centric visualizations to identify modules, investigate overlaps, and cluster pathways. This case study demonstrates their use to investigate the roles of the pathways related to established genetic associations in the context of Parkinson's disease (PD).

Pathway enrichment with Fisher's exact test using a gene panel associated with PD reviewed by Brás et al.<sup>28</sup> (the gene set will be referenced as PDgset) yielded over 300 pathways containing at least one of the panel's genes (Fig. 3a). We discarded pathways with fewer than two genes from PDgset, that were larger than 300 genes, or that were not found to be statistically significant (false discovery rate >5%) after applying multiple hypothesis testing correction with the Benjamini–Yekutieli method under dependency.<sup>29</sup>

Three views were used to assist in the interpretation of the remaining 29 enriched pathways: a pathway network view was used to identify pathway modules, a pathway overlap view was used to explore the intersections and cross-talks between pathways, and a pathway dendrogram view was used for clustering.

The pathway network view renders a pathway-to-pathway network in which nodes represent pathways and weighted edges represent their corresponding gene set similarities in a similar fashion to PathwayConnector.<sup>30</sup> For the PDgset, this visualization helped us to define six different modules (i.e., groups of pathways) by removing edges with a weight lower than 0.2 (Fig. 3b). The largest module (labeled as  $M_1$ ) contained pathways related to the processes of endocytosis and vesicle transport, both of which are putatively disrupted in PD.<sup>31</sup>  $M_2$  comprised pathways related to PTK6 signaling such as the Reactome pathway, PTK6 promotes HIF1A stabilization, whose high pathway enrichment significance ( $q$ -value = 0.0005), as well as its role in regulating another PDgset gene, ATP13A2,<sup>32</sup> suggests that it may be linked to PD. ATP13A2 is directly responsible for Kufor-Rakeb syndrome,<sup>33</sup> a rare juvenile form of PD, and participates in two other PD mechanisms: lysosomal iron storage and mitochondrial stress. Because pathways related to these two mechanisms (i.e., Lysosome pathway





**Fig. 2** **a** An Euler diagram summarizing the human gene-centric coverage of KEGG, Reactome, and WikiPathways compared to the universe of all genes from HGNC (more details in Supplementary Table 1). **b** Histogram views present gene promiscuity or pathway size distributions. **c** The pathway similarity landscape of KEGG visualized as a heatmap

from KEGG, Pink/Parkin mediated mitophagy from Reactome, and Mitophagy pathway from both KEGG and Reactome;  $M_4$ ) were also enriched by pathway enrichment analysis, we investigated the role of ATP13A2 in PD further.

ATP13A2 is activated by phosphatidylinositol(3,5)bisphosphate, a particular phosphatidylinositol involved in  $M_3$  pathways (phosphatidylinositol metabolism and signaling pathways). Because this activation leads to a reduction in mitochondrial stress and  $\alpha$ -synuclein toxicity, two hallmarks of PD, ATP13A2 has been proposed as a therapeutic target.<sup>34</sup> Ultimately, the exploration of the similarities and cross-talks between these three modules suggests further investigation of the candidate PD gene ATP13A2. Ultimately, this view complements pathway enrichment in the identification of pathway modules, exploration pathway cross-talks, and prioritization of genes for further study.

While the pathway network viewer provides an overview of the different modules and their cross-talks, it does not reveal information about their contained pathways' boundaries and intersections. Therefore, we implemented the pathway overlap view; an interactive Euler diagram that allows exploration of pathway demarcations (Fig. 3c). We employed this view to identify the set of genes common to all pathways in  $M_5$ , a module comprising the two Alzheimer's disease (AD) and two PD pathways from KEGG and WikiPathways. Subsequently, we used the ComPath pathway enrichment wizard to investigate in which pathways the common five genes identified (APAF1, CASP3, CASP9, CYCS, and SNCA) participate. The analysis revealed that they are predominantly involved in apoptosis, an important process in both AD and PD pathophysiology.<sup>35,36</sup>

The third visualization renders the results of the hierarchical clustering approach described in Chen et al. in the form of a dendrogram, enabling deterministic pathway grouping based on gene set similarity. We used this view in the PDgset example to assign the pathways without module membership to the closest module (Supplementary Figure S2). The dendrogram proposed merging three previously unassigned pathways into  $M_2$  (i.e., Allograft Rejection, MAPK Signaling pathway, and Rasp1 signaling pathway). Additionally, the resulting dendrogram from clustering revealed hierarchical relationships between pathways (e.g., Pink/Parkin Mediated Mitophagy is a subset of the Reactome Mitophagy pathway), information that can be used to establish pathway mappings, as we show in the following case study.

**Case study III: establishing mappings between pathway databases**  
ComPath, as well as other tools, have demonstrated the benefits of integrating pathway knowledge from diverse resources to improve biological functional analysis.<sup>9,10,18</sup> However, even after overcoming the technical hurdle of harmonizing different formats used by different databases, these integrative approaches must be complemented by mappings at a pathway level in order to have

cross references between databases; thus, improving their interoperability. Such information could then be used to first link related pathways and then investigate their interplays, explore the consistency of their boundaries, calculate their discrepancies and agreements, or simply contextualize the knowledge around a certain biological process.

In order to address this, ComPath introduces a curation environment in which users from the scientific community can propose and maintain a collection of established mappings between pathways from various databases. This laborious task is facilitated by the interactive visualizations (i.e., a dendrogram view and a similarity landscape heatmap) presented in the previous case studies as well as dedicated pathway pages where the content, descriptions, references, and the established mappings can be examined (Fig. 4a). Furthermore, ComPath suggests the most similar pathways based on this information so users can propose new mappings. This new mappings are included into the mapping catalog that serves as a search interface as well as a distribution platform for mappings (Fig. 4b). In addition, the mapping catalog promotes community engaging incorporating a voting system where authenticated users can agree or disagree on mappings; this way, proposed mappings with a net sum of votes  $>3$  are automatically registered as accepted.

After an exhaustive investigation of all possible mappings between pathways in KEGG, Reactome, and WikiPathways (see Methods), we identified 58 equivalencies between KEGG and Reactome, 64 between Reactome and WikiPathways, and 55 between KEGG and WikiPathways. Of these equivalent pathways, 21 are shared between the three resources (Fig. 5 and Supplementary Table 4). We also identified 247 hierarchical relationships between KEGG and Reactome, 597 between KEGG and WikiPathways, and 564 between Reactome and WikiPathways. After considering these, approximately 26% of KEGG, 70% of Reactome, and 35% of WikiPathways did not share any mappings with any other database (Supplementary Figure S4). The high uniqueness observed in Reactome could be attributed to several factors: its small pathway sizes, its high granularity, and its high coverage of HGNC (Fig. 2a).

The results of this curation effort are distributed at <https://github.com/ComPath/resources> and <https://compath.scai.fraunhofer.de/> so they can be revised, updated, and exploited by the research community hoping that this work serves as a first endeavor towards unifying pathway knowledge.

## DISCUSSION

The lack of a lingua franca in systems biology hampers the harmonization that would enable the exploration of the coverage, agreements, or discrepancies in the pathway knowledge. Harmonizing this information is an important step to better comprehend and model biology as well as improve the bioinformatics pipelines

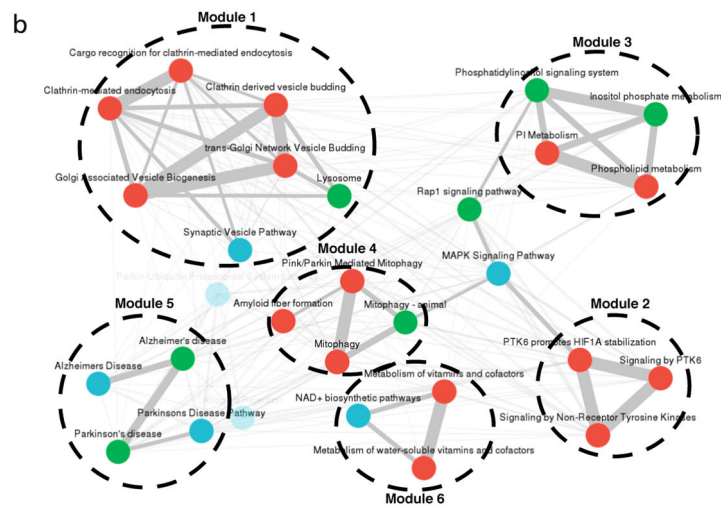
a Query Results <sup>o</sup>

|                              |  |
|------------------------------|--|
| Gene Symbols Submitted (34)  | FBXO7, ATP13A2, STX18, GAK, SYT11, RAB29, MIR4697HG, PLA2G6B, SYNJ1, GCH1, PARK7, MAPT, SCARB2, INPP5F, LRRK2, DDRGK1, MCCC1, AGMSD, VPS35, SNCA, FGF20, HLA-DRB5, PRKN, RIT2, SREBF1, PINK1, GBA, GPNMB, SIPA1L2, CDDC82, STK39, VPS13C, BST1, DNAJC8 |
| Genes not in any pathway (5) | DDRGK1, CDDC82, STK39, VPS13C, MIR4697HG   |
| Number of Pathways Mapped    | 89   |
| Select All Pathways          | <input type="checkbox"/>   |

First, select your pathways of interest and then, choose the type of analysis to perform. The "Overlap View" displays the boundaries between the selected pathways represented as Venn or Euler diagrams. The "Cluster View" renders an interactive dendrogram of the pathways clustered based on their distances. Finally, the "Network View" displays the knowledge around the selected pathways as well as the similarity between them enabling to identify interplays.

[Overlap View](#) [Cluster View](#) [Network View](#)

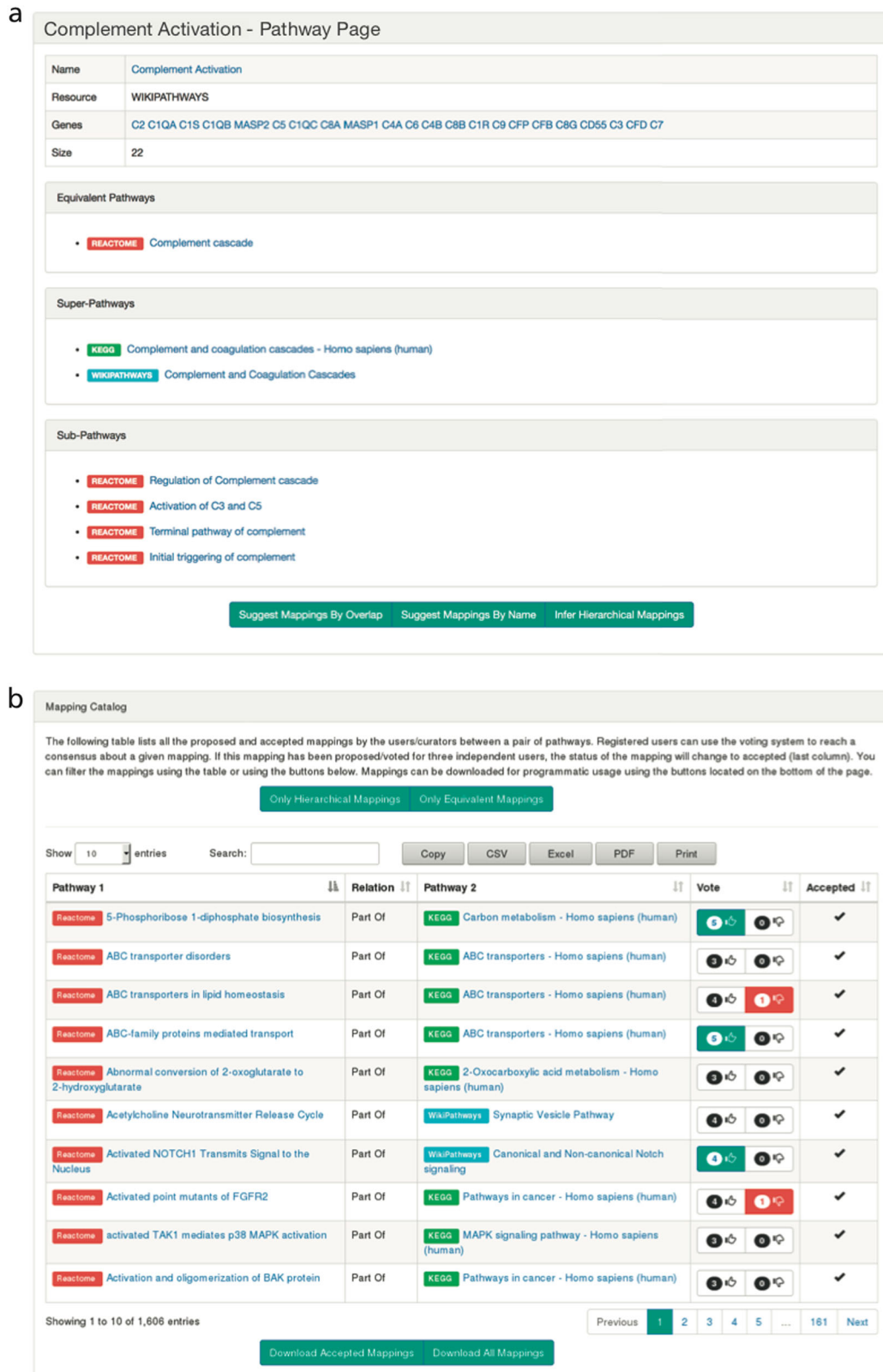
| Pathway Name  | Resource Identifier | Adjusted p-value | Genes Mapped | Pathway Size |
|---|---------------------|------------------|--------------|--------------|
| <input checked="" type="checkbox"/> Parkinsons Disease Pathway                  | WP2371              | 0.0              | 6            | 84           |
| <input checked="" type="checkbox"/> NAD+ biosynthetic pathways                  | WP3645              | 0.0035           | 2            | 22           |
| <input checked="" type="checkbox"/> Synaptic Vesicle Pathway                    | WP2267              | 0.0111           | 2            | 52           |
| <input checked="" type="checkbox"/> MAPK Signaling Pathway                      | WP382               | 0.0122           | 3            | 249          |
| <input checked="" type="checkbox"/> Parkin-Ubiquitin Proteasomal System pathway | WP2359              | 0.0143           | 2            | 70           |
| <input checked="" type="checkbox"/> Allograft Rejection                         | WP2328              | 0.0191           | 2            | 89           |



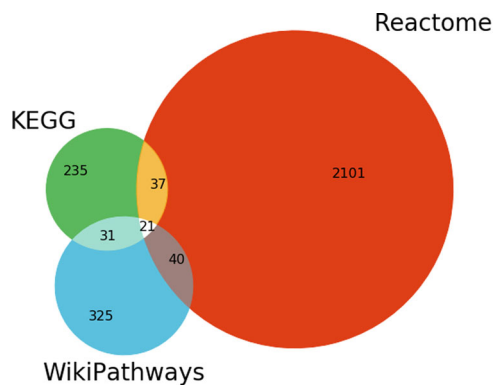
**c**



**Fig. 3** **a** Results of pathway enrichment using the PDgset as input using the ComPath pathway enrichment wizard. We would like to remark that enrichment results might change over time since ComPath regularly updates their underlying pathway databases. In order to promote reproducibility, the current version of the databases is displayed in the ComPath overview page and older versions can be provided upon request. **b** The Pathway Network Viewer displays the similarity around a selection of pathways. **c** The Pathway Overlap View depicts the overlaps and intersection of pathways enriched from the PDgset



**Fig. 4** **a** The pathway info view introduces basic pathway information such as its participating molecular entities, references, or mappings and enables automatic mapping suggestions based on different similarity metrics. Furthermore, the mappings of the selected pathway can be visualized with a dynamic view that enables exploration of multiple levels of its hierarchy (Supplementary Figure S3). **b** The mappings view allows users to browse established mappings, propose new mappings, and give feedback on putative mappings



**Fig. 5** Venn diagram illustrating the overlaps of equivalent pathways between KEGG, Reactome, WikiPathways resulting from the curation exercise. Note: the number of overlapping pathways in the Venn diagram do not exactly match the number of equivalent mappings since there are equivalent pathways within WikiPathways that, when mapped to another database, could have more than one equivalent pathway. For example, there are two equivalent Wnt signaling pathways in WikiPathways that are both mapped to their corresponding Reactome pathway. This is resolved to a unique in the Venn diagram. A list of intra-database equivalent pathways is presented in the Supplementary Table 3

that utilize this knowledge to elucidate biological insights. As a first step towards closing this gap, we have implemented an environment capable of accommodating the pathway knowledge from multiple databases in order to facilitate its exploration and analysis through a web application. The flexibility of ComPath enables the incorporation of additional databases as well as dynamic update of its resources; the latter of which is often neglected, but can have a significant effect on derived analyses.<sup>37</sup> Additionally, an embedded curation interface allow users to curate and establish mappings between pathways. Accordingly, we used ComPath to conduct extensive curation work to link the pathways from three major pathway databases in order to evaluate their similarities and differences. This mapping catalog serves as a first effort towards unifying and linking pathway information across databases that can later be adopted by the original databases or to create ontologies that store these mappings. Because databases regularly add new pathways and update gene identifiers, we plan to update ComPath biannually as well as curate mappings for these newly added pathways – current mappings do not have to be updated since the focus of the pathway does not change.

The common genes between KEGG, Reactome, and WikiPathways covered the majority of pathways, indicating that their pathway knowledge is partially biased towards this shared gene set, even while there are still thousands of genes that have not yet been functionally annotated to pathways. Furthermore, our curation effort revealed that a surprisingly low number of pathways (21) were equivalent between KEGG, Reactome, and WikiPathways. On the other hand, the number of mapped pathways increased significantly when the hierarchical mappings were considered, revealing the inconsistent granularity employed to delineate pathway boundaries.

Although the absence of topological pathway information in ComPath is an irrefutable limitation in this study, gene-centric approaches enable a reduction of complexity in pathway comparison as well as integration of resources which do not provide topology information.<sup>10</sup> Furthermore, recent studies revealed significant differences across a large sample of topology-based pathway analysis methods,<sup>38</sup> and highlighted that gene sets alone might be sufficient to detect an enriched pathway under realistic circumstances.<sup>39</sup> Hence, even if the abstraction of pathways as gene sets might not exploit all the

existing pathway information, it is sufficient to drive an investigation of the pathway knowledge.

The established inter-database mappings allowed to link pathways from three major databases, opening the door towards a better integration of the pathway knowledge. In the future, these links can be used to complement and fill pathway knowledge as well as to conduct a precise evaluation of equivalent or related pathways by exploiting the available format converters such as the converter from Reactome to WikiPathways.<sup>40</sup> Furthermore, ComPath have been designed to accommodate multiple types of molecular entities participating in pathways (i.e. Reactome chemical information); thus, enabling to replicate the analyses presented with lipid or metabolite databases such as LIPEA<sup>41</sup> or HMDB.<sup>42</sup>

In summary, we demonstrated that ComPath serves as an exploratory, analytic, and curation framework for pathway databases. Furthermore, we showed how the ComPath web application can complement enrichment approaches to elucidate and prioritize pathways and genes related to interesting biological phenomenon. Finally, we hope that the implementation of a curation ecosystem and the first mapping efforts conducted in this work pave the way towards unifying the pathway knowledge.

## METHODS

### ComPath framework

At its core, ComPath is a framework for integrating pathway and gene set databases. We defined a set of guidelines for implementing wrappers around the processes of downloading data, transforming it into a common data model, and making queries. These guidelines are encoded in an abstract class with the Python programming language such that new plugins can be quickly implemented for new resources. Each implementation must have a mapping between genes and pathways as well as functions for exporting pathways as gene sets, performing pathway enrichment analysis, and performing reasoning/inference over pathway hierarchies.

### Compath plugins

We implemented plugins for four major public pathway databases: KEGG, Reactome, WikiPathways, and MSigDB.<sup>1–4</sup> They can be used individually as a way of extracting, updating, and exploring the pathways contained within the database. Additionally, they can be used jointly in the ComPath web application where the pathways from multiple databases are integrated for their exploration, analysis, and curation.

### ComPath web application

The web application was implemented in the Python programming language using the Flask microframework and a suite of its extensions. The compatibility between Flask and the data models defined in all pathway plugins allows the integration and harmonization of the pathway knowledge in an extensible manner. To illustrate the flexibility of ComPath, we have included plugins for the Alzheimer's disease and Parkinson's disease gene sets associated with disease-specific mechanisms from NeuroMMSig<sup>43</sup> in the public version of the ComPath web (<https://compath.scai.fraunhofer.de/>).

ComPath leverages a variety of state-of-the-art libraries for visualization and exploration of pathway knowledge. We chose Bootstrap for the design of the website since its responsive design retains full compatibility across all devices. Interactive visualizations are generated using several Javascript libraries, including D3.js, Clustergrammer.js,<sup>24</sup> and Cytoscape.js.<sup>44</sup>

We implemented a RESTful API documented with an OpenAPI specification that can be accessed through the ComPath instance released at <https://compath.scai.fraunhofer.de/apidocs>. The API enables users to programmatically extract mapping information and perform queries using different genes or pathways identifiers.

### Code availability

The source code for ComPath and its plugins can be found on GitHub (<https://github.com/ComPath> and <https://github.com/Bio2BEL>) under the MIT license. Both the plugins and the web application can be installed with

PyPI (<https://pypi.org>), the main packaging system for Python. Furthermore, we have included a Dockerfile to enable reproducing the ComPath environment with Docker (<https://www.docker.com/>). Finally, documentation is included in each GitHub repository and it is also accessible at Read the Docs (<https://readthedocs.org>).

### Estimating pathway similarity

While a variety of indices (e.g., Jaccard, Sørensen–Dice, Tversky) have been used to assess the similarity between sets, the Szymkiewicz–Simpson coefficient (Eq. 1) is most appropriate for comparing sets widely varying in size. Similarly to previous studies, we have chosen this index to not only calculate pathway similarity but also reveal contained pathways (i.e., when most of the nodes from a small pathway are in a larger pathway) to indicate potential hierarchical relationships.<sup>10,45–47</sup>

$$S_{(X,Y)} = \frac{|X \cap Y|}{\min(|X|, |Y|)}$$

Equation 1. The Szymkiewicz–Simpson coefficient calculates the similarity between two sets (X and Y) where  $0 \leq S \leq 1$ . The similarity is the size of the intersection of the two sets divided by the size of the smaller.

### Curation of pathway mappings

Here, we describe a semi-automatic curation procedure we used in order to systematically generate equivalency and hierarchical mappings between the human pathways originating from KEGG, Reactome, and WikiPathways. Here, it is important to note that we have only focused on generating mappings for the pathways originating from each of the three resources, not their imported pathways from other databases (e.g., WikiPathways imported Reactome pathways that are evidently equivalent to the ones in Reactome). First, we define two types of mappings:

1. *equivalentTo*. An undirected relationship denoting both pathways refer to the same biological process. The requirements for this relationship are:
  - Scope: both pathways represent the same biological pathway information.
  - Similarity: both pathways must share at minimum of one overlapping gene.
  - Context: both pathways should take place in the same context (e.g., cell line, physiology).
2. *isPartOf*. A directed relationship denoting the hierarchical relationship between the pathway 1 (child) and 2 (parent). The requirements are:
  - Subset scope: the subject (pathway 1) is a subset of pathway 2 (e.g., reactome pathway hierarchy).
  - Similarity: same as above.
  - Context: same as above.

We generated all possible mappings between pathways in each database (KEGG–WikiPathways, KEGG–Reactome, and WikiPathways–Reactome) and prioritized them based on the follow two independent metrics that have been proposed to calculate pathway similarity:<sup>10</sup>

1. Lexical similarity between each pair of pathways' names was calculated using the Levenshtein distance.<sup>48</sup>
2. Content similarity between each pair of pathways' genes was calculated using the previously described Szymkiewicz–Simpson coefficient.

After prioritization, our three curators from different areas of expertise (neuroscience, medicine, and biology) independently evaluated both similarities and the scope and context included in the pathway descriptions to assign the mapping types and to remove false positives. Furthermore, we investigated possible intra-database mappings within KEGG and WikiPathways since these resources do not yet contain hierarchical relationships. Finally, our curators combined the results and re-evaluated them to generate a consensus mapping file. It is available at <https://github.com/ComPath/resources> under the MIT License.

### ACKNOWLEDGEMENTS

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

### AUTHOR CONTRIBUTIONS

M.H.A. and D.D.F. conceived and designed the study. D.D.F. implemented ComPath and the pathway database plugins with help from C.T.H. D.D.F., C.B.A., and J.M.L. curated the pathway mappings. D.D.F. and C.T.H. wrote the paper. M.H.A. reviewed the content.

### ADDITIONAL INFORMATION

**Supplementary information** accompanies the paper on the *npj Systems Biology and Applications* website (<https://doi.org/10.1038/s41540-018-0078-8>).

**Competing interests:** The authors declare no competing interests.

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

### REFERENCES

1. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45** (D1), D353–D361 (2016).
2. Fabregat, A. et al. The reactome pathway knowledgebase. *Nucleic Acids Res.* **46** (D1), D649–D655 (2017).
3. Slenter, D. N. et al. WikiPathways: A multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res.* **46**(D1), D661–D667 (2017).
4. Liberzon, A. et al. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
5. Krämer, A., Green, J., Pollard, J. Jr & Tugendreich, S. Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics* **30**, 523–530 (2013).
6. Kutmon, M. et al. WikiPathways: Capturing the full diversity of pathway knowledge. *Nucleic Acids Res.* **44**(D1), D488–D494 (2015).
7. Demir, E. et al. The BioPAX community standard for pathway data sharing. *Nat. Biotechnol.* **28**, 935 (2010).
8. Hucka, M. et al. The systems biology markup language (SBML): A medium for representation and exchange of biochemical network models. *Bioinformatics* **19**, 524–531 (2003).
9. Cerami, E. G. et al. Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res.* **39**, D685–D690 (2010).
10. Belinky, F., et al. PathCards: Multi-source consolidation of human biological pathways. *Database*, 2015, bav006 (2015).
11. Yamada, T. et al. iPath2.0: Interactive pathway explorer. *Nucleic Acids Res.* **39** (suppl\_2), W412–W415 (2011).
12. Perfetto, L. et al. SIGNOR: A database of causal relationships between biological entities. *Nucleic Acids Res.* **44**(D1), D548–D554 (2015).
13. Petri, V. et al. The pathway ontology–updates and applications. *J. Biomed. Semantics*. **5**, 7 (2014).
14. Iyappan, A. et al. Towards a pathway inventory of the human brain for modeling disease mechanisms underlying neurodegeneration. *J. Alzheimer's. Dis.* **52**, 1343–1360 (2016).
15. Doderer, M. S. et al. Pathway Distiller–multisource biological pathway consolidation. *BMC Genomics* **13**, S18 (2012).
16. Daniels, K., and Giraud-Carrier, C. Learning the threshold in hierarchical agglomerative clustering. In *5th International Conference on Machine Learning and Applications*, 2006. ICMLA'06. (pp. 270–278). IEEE (2006).
17. Khatri, P., Sirota, M. & Butte, A. J. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.* **8**, e1002375 (2012).
18. Kuleshov, M. V. et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**(W1), W90–W97 (2016).
19. Cary, M. P., Bader, G. D. & Sander, C. Pathway information for systems biology. *FEBS Lett.* **579**, 1815–1820 (2005).
20. Subramanian et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
21. Bauer-Mehren, A., Furlong, L. I. & Sanz, F. Pathway databases and tools for their exploitation: Benefits, current limitations and challenges. *Mol. Syst. Biol.* **5**, 290 (2009).

22. Chowdhury, S. & Sarkar, R. R. Comparison of human cell signaling pathway databases—evolution, drawbacks and challenges. *Database* **2015**, bau126 (2015).
23. Stobbe, M. D., Houten, S. M., Jansen, G. A., van Kampen, A. H. & Moerland, P. D. Critical assessment of human metabolic pathway databases: a stepping stone for future integration. *BMC Syst. Biol.* **5**, 165 (2011).
24. Fernández, N. F. et al. Clustergrammer, a web-based heatmap visualization and analysis tool for high-dimensional biological data. *Sci. Data* **4**, 170151 (2017).
25. Reimand, J. et al. g: Profiler—a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res.* **44**(W1), W83–W89 (2016).
26. Pathan, M. et al. FunRich: an open access standalone functional enrichment and interaction network analysis tool. *Proteomics* **15.15**, 2597–2601 (2015).
27. Huang, W. et al. The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol.* **8**, R183 (2007).
28. Brás, J., Guerreiro, R. & Hardy, J. Snapshot: genetics of Parkinson's disease. *Cell* **160**, 570–570 (2015).
29. Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* **29**, 1165–1188 (2001).
30. Minadakis, G., et al. PathwayConnector: Finding complementary pathways to enhance functional analysis, *Bioinformatics*, 10.1093/bioinformatics/bty693 (2018).
31. Perrett, R. M., Alexopoulou, Z. & Tofaris, G. K. The endosomal pathway in Parkinson's disease. *Mol. Cell. Neurosci.* **66**, 21–28 (2015).
32. Rajagopalan, S., Rane, A., Chinta, S. J. & Andersen, J. K. Regulation of ATP13A2 via PHD2-HIF1 $\alpha$  signaling is critical for cellular iron homeostasis: implications for Parkinson's disease. *J. Neurosci.* **36**, 1086–1095 (2016).
33. Gusdon, A. M., Zhu, J., Van Houten, B. & Chu, C. T. ATP13A2 regulates mitochondrial bioenergetics through macroautophagy. *Neurobiol. Dis.* **45**, 962–972 (2012).
34. Holemans, T. et al. A lipid switch unlocks Parkinson's disease-associated ATP13A2. *Proc. Natl Acad. Sci. USA* **112**, 9040–9045 (2015).
35. Obulesu, M. & Lakshmi, M. J. Apoptosis in Alzheimer's disease: An understanding of the physiology, pathology and therapeutic avenues. *Neurochem. Res.* **39**, 2301–2312 (2014).
36. Tatton, W. G., Chalmers-Redman, R., Brown, D. & Tatton, N. Apoptosis in Parkinson's disease: signals for neuronal degradation. *Ann. Neurol.* **53**(S3), S61–70, [https://doi.org/10.1002/\(ISSN\)1531-8249](https://doi.org/10.1002/(ISSN)1531-8249) (2003).
37. Wadi, L. et al. Impact of outdated gene annotations on pathway enrichment analysis. *Nat. Methods* **13**, 705 (2016).
38. Ilnatova, I., Popovici, V. & Budinska, E. A critical comparison of topology-based pathway analysis methods. *PLoS One* **13**, e0191154 (2018).
39. Bayerlová, M. et al. Comparative study on gene set and pathway topology-based enrichment methods. *BMC Bioinformatics* **16**, 334 (2015).
40. Bohler, A. et al. Reactome from a WikiPathways perspective. *PLoS Comput. Biol.* **12**, e1004941 (2016).
41. Acevedo, A., Duran, C., Ciucci, S., Gerl, M., and Cannistraci, C. V. LIPEA: Lipid Pathway Enrichment Analysis. bioRxiv, <https://doi.org/10.1101/274969> (2018).
42. Wishart, D. S. et al. HMDB 4.0: The human metabolome database for 2018. *Nucleic Acids Res.* **46**(D1), D608–D617 (2017).
43. Domingo-Fernández, D. et al. Multimodal mechanistic signatures for neurodegenerative diseases (NeuroMMSig): a web server for mechanism enrichment. *Bioinformatics* **33**, 3679–3681 (2017).
44. Franz, M. et al. Cytoscape.js: A graph theory library for visualisation and analysis. *Bioinformatics* **32**, 309–311 (2015).
45. Chen, Y. A. et al. Integrated pathway clusters with coherent biological themes for target prioritisation. *PLoS One* **9**, e99030 (2014).
46. Pita-Juarez, Y. et al. The pathway coexpression network: Revealing pathway relationships. *PLoS Comput. Biol.* **14**, e1006042 (2018).
47. Katiyar, A., Sharma, S., Singh, T. P. & Kaur, P. Identification of shared molecular signatures indicate the susceptibility of endometriosis to multiple sclerosis. *Front. Genet.* **9**, 42 (2018).
48. Levenshtein, V. I. Binary codes capable of correcting deletions, insertions, and reversals. *Sov. Phys. Dokl.* **10**, 707–710 (1966).



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

© The Author(s) 2018

## Conclusions

We have developed ComPath, a flexible framework for harmonizing and integrating key molecular players such as gene and chemical sets curated in pathway databases. ComPath is complemented with a web application [119] that accommodates heterogeneous data to enable the exploration, analysis, and curation of pathway knowledge. The web application offers i) analytical functionalities to conduct pathway enrichment, ii) novel visualizations to investigate overlap and crosstalk, both at the database and pathway level, and iii) a community-driven interface to curate inter-database pathway mappings. We illustrate the utility of ComPath by analyzing pathway crosstalk using PD hallmark genes and by curating the first mappings across three of the major pathway databases (i.e., KEGG, Reactome, and WikiPathways) [17, 20, 118].

While related pathways can be found in disparate resources, the biomedical community can greatly benefit from a pathway catalogue which establishes the degree to which pathways in one resource are related to those in another. The dataset curated in this work fills this void by connecting three of the major pathway databases and enables systematic approaches to compare and investigate similarity across specific pathways or databases. Finally, it provides the first overview of the diversity in the pathway representations captured across the three databases. Thus, highlighting the importance of connecting pathway resources to provide a global and comprehensive picture of pathway knowledge.





# 3 PathMe: Merging and exploring mechanistic pathway knowledge

## Introduction

In parallel to the development of pathway databases during the last decades, numerous formats have been proposed to formalize pathway knowledge. This variety of formats enables curators to represent pathways optimally for their specific research questions. However, this presents a major roadblock to integrative approaches. Though frameworks such as Pathway Commons [25], graphite [63], and OmniPath [62] have integrated information from multiple pathway databases, they do not harmonize their heterogeneity in multi-scale biological entities and relationships. Accommodating this complementary information into a common schema is instrumental to provide a comprehensive view of the pathway landscape. This becomes even more evident when existing databases contain largely non-overlapping set of pathways, as we have shown in the previous chapter. This challenge prompted us to develop PathMe, a sophisticated software that harmonizes pathway knowledge using BEL as an overarching and integrative schema.


Reprinted with permission from "Daniel Domingo-Fernández *et al.*. PathMe: Merging and exploring mechanistic pathway knowledge. *BMC Bioinformatics*, 20:243". Copyright © Daniel Domingo-Fernández 2019.

SOFTWARE

Open Access



# PathMe: merging and exploring mechanistic pathway knowledge

Daniel Domingo-Fernández<sup>1,2\*</sup> , Sarah Mubeen<sup>1,2</sup>, Josep Marín-Llaó<sup>1</sup>, Charles Tapley Hoyt<sup>1,2</sup> and Martin Hofmann-Apitius<sup>1,2</sup>

## Abstract

**Background:** The complexity of representing biological systems is compounded by an ever-expanding body of knowledge emerging from multi-omics experiments. A number of pathway databases have facilitated pathway-centric approaches that assist in the interpretation of molecular signatures yielded by these experiments. However, the lack of interoperability between pathway databases has hindered the ability to harmonize these resources and to exploit their consolidated knowledge. Such a unification of pathway knowledge is imperative in enhancing the comprehension and modeling of biological abstractions.

**Results:** Here, we present PathMe, a Python package that transforms pathway knowledge from three major pathway databases into a unified abstraction using Biological Expression Language as the pivotal, integrative schema. PathMe is complemented by a novel web application (freely available at <https://pathme.scai.fraunhofer.de/>) which allows users to comprehensively explore pathway crosstalk and compare areas of consensus and discrepancies.

**Conclusions:** This work has harmonized three major pathway databases and transformed them into a unified schema in order to gain a holistic picture of pathway knowledge. We demonstrate the utility of the PathMe framework in: i) integrating pathway landscapes at the database level, ii) comparing the degree of consensus at the pathway level, and iii) exploring pathway crosstalk and investigating consensus at the molecular level.

**Keywords:** Bioinformatics, Pathways, Database integration, Network analysis, Biological networks, Biological expression language

## Background

The interpretations of molecular signatures that are typically yielded by genome-scale experiments are often supported by pathway-centric approaches through which mechanistic insights can be gained by pointing at a set of biological processes. Thus, parallel to the development of novel data-driven approaches, pathway databases emerged as comprehensive resources that could be used to complement analyses with prior knowledge. These resources have embraced standard file formats and schemata in order to facilitate the exchange of pathway knowledge. However, each resource has chosen a different one and though these formats possess overlapping capabilities to produce computational models of biology, their intended purposes and

applications are somewhat distinct. For instance, Systems Biology Markup Language (SBML) is a standard for the representation of computational models of systems biology, Systems Biology Graphical Notation (SBGN) facilitates the storage and exchange of signaling pathways, metabolic networks and gene regulatory network information, and Biological Pathway Exchange (BioPAX) has been designed with the purpose of establishing a common exchange format for biological pathway data [12, 25, 31]. A variety of formats offer the scientific community multiple approaches to curate pathway knowledge. However, a multitude of diverse formats and a lack of interoperability between them tends to hamper efforts to collate the knowledge contained in pathway databases. In practice, this has led to the generation of data silos derived from the gradual detachment of complementary work from different research groups which use distinct modeling languages. Therefore, metadatabases such as Pathway Commons [9] and ConsensusPathDB [26], which incorporate

\* Correspondence: [daniel.domingo.fernandez@scai.fraunhofer.de](mailto:daniel.domingo.fernandez@scai.fraunhofer.de)

<sup>1</sup>Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing, 53754 Sankt Augustin, Germany

<sup>2</sup>Bonn-Aachen International Center for IT, Rheinische Friedrich-Wilhelms-Universität Bonn, 53115 Bonn, Germany



© The Author(s). 2019 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated.

several different primary resources in their data warehouses, and integrative software applications such as *graphite* [37] and OmniPath [41] have been created with the primary intention to integrate pathway knowledge from multiple databases. Beyond these, other approaches such as those taken in PathCards [4], RaMP [45], and ComPath [14], have focused on integrating gene sets and chemical knowledge related to pathways, but without including their topological information (i.e., relationships were excluded from the network). For instance, ComPath, the precursor of this work, harmonized pathway information at the gene level in order to conduct extensive manual curation that mapped and cross-referenced pathway representations across databases. This mapping catalog reveals which pathways are covered by which database (e.g., pathway in resource X is equivalent to pathway in resource Y) and facilitates comparing the results of pathway enrichment methods.

The representation of pathway knowledge can span several scales including molecular events, cellular processes and/or phenotypes, which are captured in varying degrees by integrative resources. For example, ConsensusPathDB and *graphite* effectively account for and harmonize metabolites, genes, and proteins when integrating pathways from multiple databases, but exclude biological types at higher order scales, such as biological processes, and other entities, such as miRNAs. On the other hand, Pathway Commons can incorporate multiple scales of biology by retaining original entity identifiers; however, it does not directly harmonize biological entities.

An ongoing challenge in harmonizing pathway resources is the use of distinct nomenclatures by individual databases. For example, for gene and gene products there exist several standard terminologies such as ENTREZ [32], UniProt [2], Ensembl [24], and HGNC [35], or for chemicals, ChEBI [21], ChEMBL [18], and PubChem [6]. Despite the availability of standard terminologies, some resources still assign biological entities and concepts to internal database identifiers. Therefore, mappers are necessary to normalize identifiers and facilitate resource harmonization (van [42]). Similarly, the harmonization of biological relationships is required to unify heterogeneous networks. While several format translators can convert interactions across formats [5, 7, 13, 20, 44], the process of harmonizing relationships, or edges in pathway networks, is not trivial; thus, hampering an integrative approach comprising several databases.

Just as pathway databases should be regularly updated to incorporate continual changes in pathway definitions, pathway metadatabases should also be updated in parallel to reflect such changes; it has been shown that by using outdated resources, results of studies are strongly influenced, and follow-up studies are negatively impacted [43]. Correspondingly, approaches to harmonize pathway data

also require these considerations or they too would be subject to similar liabilities. Moreover, pathway analysis software have been recently complemented with user-friendly exploratory tools and applications such as Pathway Commons, PathVisio [29], Cytoscape.js [17], or NDEx [36], which have been specifically designed for the visualization of individuals pathways and biological networks, including at a finer, more granular level. While the scientific community has greatly benefited from the development of these tools, there is still the need to develop applications that focus on visualizing the consensus and crosstalk between multiple, disparate pathway representations. While previously mentioned attempts have succeeded in accumulating and increasing the availability of database content, there has not yet been a systematic evaluation that investigates the degree of overlap or the amount of agreements/discrepancies in related or equivalent pathways from different databases. Previous comprehensive comparisons of database content were restricted to single or small sets of pathways because of the considerable amount of manual intervention (e.g., entity/relationship normalization, image reconstruction, etc.) required to shed light on the degree of overlap of equivalent pathways [11, 39]. Conversely, conducting a systematic comparison requires harmonization of entities and biological interactions across databases and minimizing pathway information loss whilst accommodating databases into an interoperable schema (i.e., retain most of the different biological abstractions that each database offers in the transformation process). Finally, connecting and integrating pathway knowledge can enhance pathway enrichment analyses, as has already been demonstrated in a more simplistic approach by Minadakis et al., as well as drive curation and new experimentation by highlighting the consensus, discrepancies, and unexplored areas of the pathway landscape.

Here, we introduce PathMe, an extensible package that harmonizes multiple databases using Biological Expression Language (BEL) as a common interoperable schema and enables pathway knowledge evaluation and exploration powered by a stand-alone web application with a special focus on highlighting pathway crosstalks and consensus.

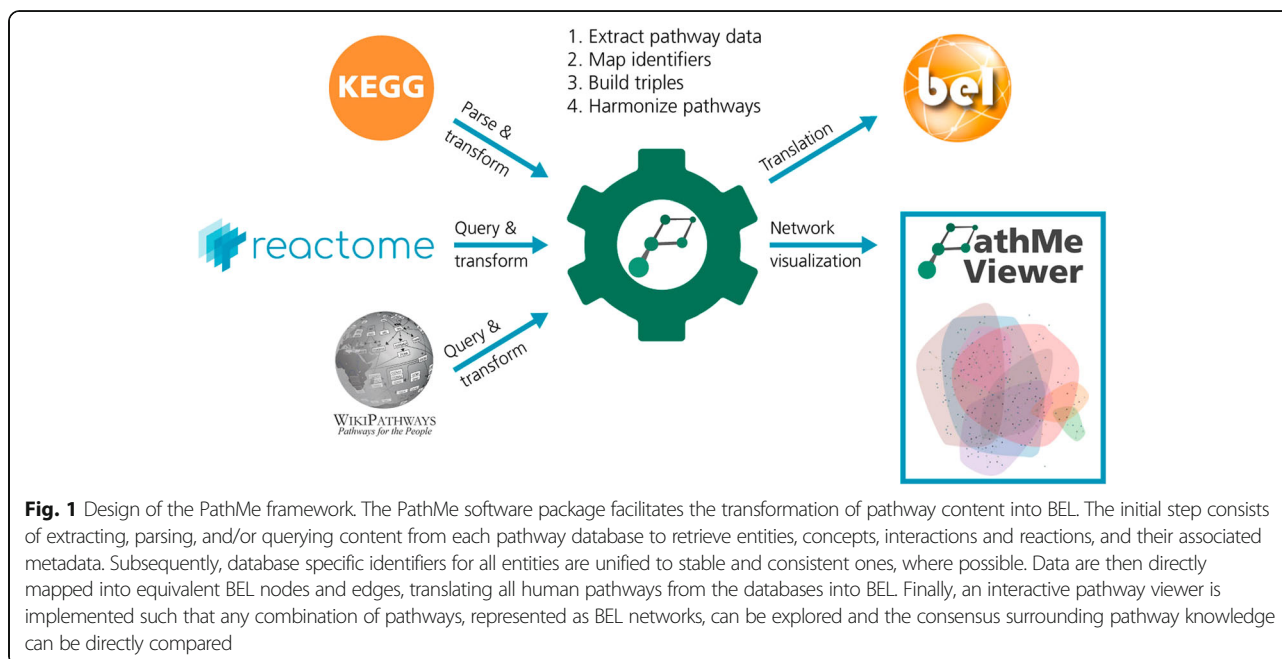
## Implementation

PathMe framework is comprised of two parts: the open-source Python package that converts the different database formats into BEL and the web application that allows for the exploration of the resulting networks (Fig. 1).

### The PathMe Python package

#### *Integrating knowledge across pathway databases*

Integrating pathway knowledge from multiple databases first requires transforming the content of each database into a common underlying schema. While multiple triple-based formats can be used to formalize pathways



in system biology, we adopted BEL as the pivotal unifying schema since it provides a reasonable trade-off between expressivity and standardized organization. Until now, we have implemented parsers for three major databases (i.e., KEGG, Reactome, and WikiPathways [15, 27, 38]) that extract pathway information and serialize it to BEL. As the principal goals of PathMe are to enable direct comparisons and explorations of pathways from different databases, cross-database mappings of identifiers and relation types are required. Accordingly, the parsers harmonize molecular entities to identifiers from standard nomenclatures as well as interaction types into their corresponding BEL relationships.

In order to harmonize entities, we prioritized standard nomenclatures for each of the modalities (e.g., genes, proteins, metabolites, etc.) included in the three studied databases (Additional file 1: Tables S1, S4, and S6). HGNC was the top-level priority namespace for genes and gene products [35]. HGNC was selected as it is recognized as an authority for standard nomenclature assignments and annotations for human genes and because the software is primarily concerned with converting human pathways. In the absence of HGNC identifiers, lower level priority namespaces were used to derive the top level HGNC identifier assignment. For instance, we aimed to use intermediate level UniProt identifiers [2] to map back to HGNC identifiers. If mappings to the prioritized namespaces were not available, genes and gene products retained their database-specific identifiers and were assigned to namespaces designated by their respective databases in order to maximize the retrieval of entities from each resource. Similarly, metabolites were prioritized to preferentially obtain ChEBI

identifiers because of ChEBI's wide usage as a source of manually curated stable identifiers and annotations for small chemical compounds [21]. In their absence, either PubChem identifiers were assigned or, once again, they retained their database-specific identifiers. Once entities were assigned to standardized identifiers, the modalities defined by the source databases were mapped to their corresponding BEL node classes (e.g., gene, protein, metabolite, biological process, etc.). Efforts were made to accommodate entities not readily mappable to BEL nodes by using BEL node classes which can incorporate flexibility in their definitions. For instance, unspecified physical entities in WikiPathways are given the abstract class label, 'DataNode'; these 'DataNodes' were mapped to BEL abundances, a category that represents the abundance of a biological entity such as a chemical or an unspecified molecule. Entity class mappings from the source databases to BEL are summarized in Additional file 1: Tables S2, S5, and S7.

Similar to the normalization of biological entities into a standardized nomenclature and their translation into corresponding BEL entity classes, distinct relationships utilized in the biological networks of different databases must too be normalized. While the versatility of BEL permitted the successful transformation of all relationships from Reactome and WikiPathways, four KEGG relationships (i.e., hidden compound, state change, dissociation, and missing interaction) could not be translated into BEL due to the lack of correspondingly equivalent edges in the BEL syntax. However, these four relationships represent non-causal interactions between biological entities and are also minimally utilized by KEGG curators. Mappings between edges from the

source databases to BEL are reported in Additional file 1: Tables S3, S5 and, S7.

#### Implementation details

PathMe relies on the individual parsers that convert the original formats from the databases to BEL. Each parser is implemented using libraries that enable the manipulation and transformation of its corresponding schemata (i.e., RDFLib for Resource Description Framework (RDF) and the xml Python package for Extensible Markup Language (XML)). Moreover, the parsers are structured into their own packages inside the main Python module to facilitate the inclusion of additional database parsers in the future. During the entity normalization process, mappings across identifiers are facilitated through the numerous packages included in the Bio2BEL framework (<https://github.com/bio2bel>) (Additional file 1: Table S9). After the normalization, entities and their relationships in each pathway are translated to BEL using the internal domain specific language (DSL) and the *BELGraph* class of the PyBEL Python software package [22]. PathMe benefits from the numerous modules implemented in the PyBEL ecosystem since it offers a variety of functionalities and algorithms that enable querying, transforming, and analyzing biological networks, as well as an export module that can output multiple formats. Finally, PathMe is distributed as a Python package through Python Package Index (PyPI) and its source code is available in GitHub at <https://github.com/PathwayMerger/PathMe>.

#### PathMe viewer

##### A web application to explore pathway knowledge

As discussed in the introduction, several visualization tools have focused on the exploration of biological networks, but have never attempted to study or evaluate the coverage, consensus, and crosstalks across heterogeneous networks. Since the particular use case of this work called for customized solutions (e.g., delineating boundaries or highlighting agreements when multiple pathways are being visualized), we also implemented a novel tool called PathMe Viewer to fulfill these unmet needs and complement the PathMe package. Since the target audience for this application are pathway curators and researchers, we opted to implement the viewer in the form of a user-friendly web application compatible with any device. The front-end extends the visualizations from BEL Commons [23] and provides an intuitive and interactive interface for visualizing and exploring the knowledge comprised in the pathway landscape. Moreover, the web application is complemented with analysis modules and network algorithms to query pathways or calculate their similarity as well as exporting options to multiple standard formats such as BEL, GraphML, or JSON so networks can be used in other software designed for visualization purposes such as Cytoscape

[17] or advanced algorithmic analyses such as SPIA [40]. Finally, the network visualization is a stand-alone component within the web application and it remains agnostic to BEL by rendering the graphics using the Node-Link JSON data format, a standard format used by popular visualization libraries; thus, facilitating the reusability of the component out of the BEL community.

#### Implementation details

PathMe Viewer follows a model-view-controller (MVC) software architecture. While the back-end is implemented in Python using the Flask microframework, the front-end is implemented in JavaScript using libraries such as jQuery (<https://jquery.com>), D3.js (<https://d3js.org>), and Bootstrap (<https://getbootstrap.com>). The source code is available at <https://github.com/PathwayMerger/PathMe-Viewer> so that all visualizations and components can be reused or extended by future applications. Furthermore, the web application is distributed through PyPI and can also be deployed with Docker which facilitates the reproducibility of this work since Docker's automated deployment process ensures that every single instance runs with the exact same settings, regardless of the host machine. Documentation is included in the GitHub repository and it is also accessible through Read the Docs (<https://pathme-viewer.readthedocs.io/en/latest/>). Finally, we provide access to a public deployment of the PathMe Viewer at <https://pathme.scai.fraunhofer.de>.

#### Calculating pathway similarity

As an application of the software, we conducted the following protocol to evaluate the degree of overlap between the three representations of each equivalent pathway (*Case scenario II*). We used a variation of the Szymkiewicz–Simpson/Overlap coefficient (Eq. 1), calculated for common molecular nodes shared between the networks. To calculate a pathway similarity index, we summed the three coefficients obtained for each individual pairwise comparison and divided this number by three to normalize to a zero-to-one scale. In other words, each pathway similarity index corresponds to a normalized sum of the individual overlaps between: i) the KEGG and Reactome representation, ii) the KEGG and WikiPathways representations, and iii) the Reactome and WikiPathways representations. Therefore, the pathway similarity index (*S*) lies between  $0 \leq S \leq 1$  (with 0 corresponding to no overlap between any of the three sets, and 1 corresponding to three fully overlapping sets).

$$S(X, Y) = \frac{|X \cap Y|}{\min(|X|, |Y|)} \quad (1)$$

Equation 1 The Szymkiewicz–Simpson coefficient calculates the similarity between two sets (*X* and *Y*) where

$0 \leq S \leq 1$ . The similarity is the size of the intersection of the two sets divided by the size of the smaller set. In this case, the sets correspond to the number of individual molecular entities excluding group nodes in the BEL graph, this is discussed in detail in the Additional file 1.

## Results

In the first two sections, we present the main functionalities of the PathMe software and web application respectively, while the following section outlines the architecture and design of the framework. Next, three case scenarios applied at increasingly granular scales of pathway knowledge are presented to illustrate the usability of the framework in database integration from a global, database-wide perspective to a detailed, path way level one.

### PathMe functions

The PathMe package offers a set of functionalities for the set of databases incorporated thus far: i) download the raw pathway files, ii) generate BEL networks and export them as binary data, iii) summarize the transformed content, and iv) calculate detailed network statistics (e.g., number of nodes, edges and their types) (Table 1). Moreover, database specific features include functionalities to flatten all group nodes (e.g., protein complexes, gene families, etc.) in KEGG and exclusively parse canonical pathways from WikiPathways and Reactome. In conclusion, these functionalities combined with the ones already offered by the PyBEL ecosystem assist bioinformaticians in transforming, exploring, and analyzing the generated pathway networks.

### PathMe viewer

Beyond the software concerned with the integration of pathway knowledge, a novel web application (i.e. PathMe Viewer) was implemented for intuitive querying, browsing, and navigating of the normalized BEL networks. Queries can be submitted for a single or a set of pathways on the main page of the viewer, as illustrated in Fig. 2a. The result of the query leads to a visualization, as seen in Fig. 2b, that renders the corresponding network.

The PathMe Viewer is powered by multiple, built-in functionalities enabling users to navigate through the pathway(s). Although the initial network layout is defined by the D3 force algorithm which enables users to

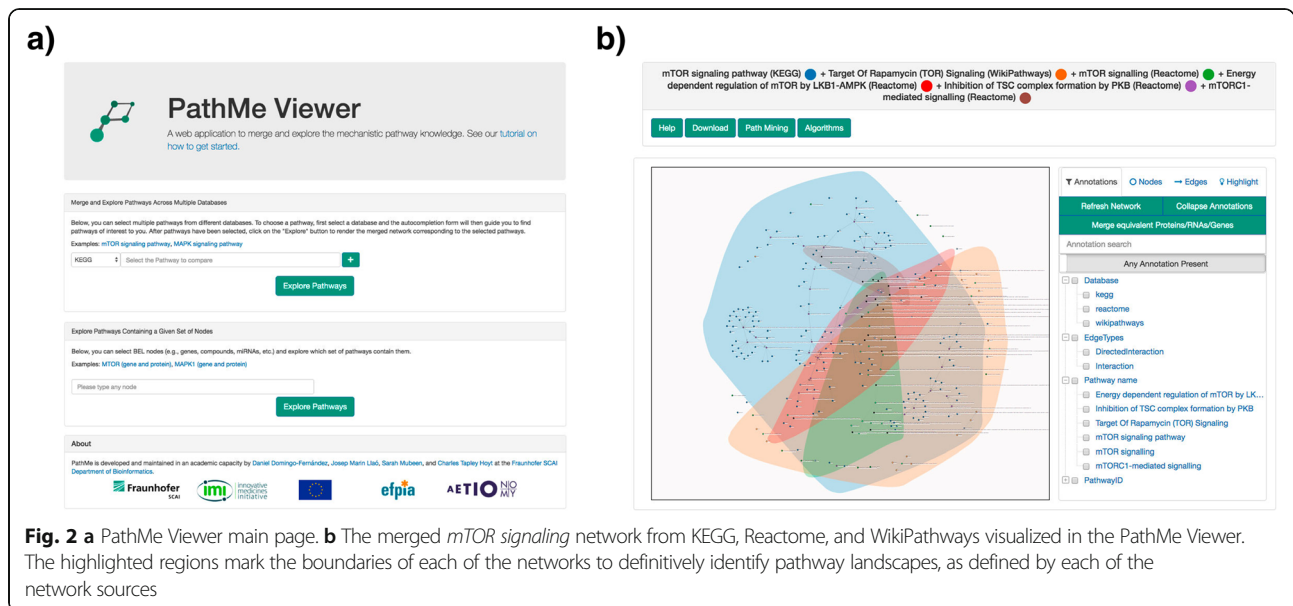
get a comprehensive overview of the relevant parts of the network, the network arrangement can also be customized by dragging and moving nodes around the viewer. Furthermore, node and edge meta-information can be accessed via double click. For nodes, this includes specifications on their name, function and namespace, while for edges, the pathway name, identifier and source database are provided. When multiple pathways are queried, marked boundaries delineate the topological landscape of each of the networks which synergistically contribute to the consolidated one to facilitate the exploration of pathway crosstalks (i.e. the interaction of pathways through their sharing of common entities) (Fig. 2b). Furthermore, search and mining tools enable navigation of the resulting network such as selecting and filtering nodes/edges or calculating paths. Another novel feature of the viewer is the automatic identification of contradictory and consensus relationships across pathways (i.e., edges between identical nodes with equivalent or opposite relationship), which are highlighted in blue/red in the network. The viewer also incorporates a functionality which collapses all BEL proteins, RNA species and genes into gene nodes. This function was included in the viewer because of the interchangeable usage of these entities by the various databases which would both preclude the ability to fairly establish if there is overlap in the network topology and to conduct fair comparisons. Finally, network algorithms such as betweenness centrality can be used to quickly identify central nodes in the network or to calculate pathway similarity as we will present in the case scenario.

### Software development techniques

Successful contributions to the bioinformatics domain are predicated by their ability to be replicated and reused. In line with community standards designed to foster these attributes, the PathMe and PathMe-Viewer packages use git (<https://git-scm.com>) for version control on GitHub (<https://github.com>), flake8 (<https://github.com/PyCQA/flake8>) to enforce code quality, setuptools (<https://github.com/pypa/setuptools>) to build distributions, pyroma (<https://github.com/regebro/pyroma>) to enforce package metadata standards, sphinx (<https://github.com/sphinx-doc/sphinx>) to build documentation, Read the Docs (<https://readthedocs.org>) to

**Table 1** Core functions of the PathMe Python package

| Function          | Description  |
|-------------------|--|
| <i>Download</i>   | Downloads the pathway files from the original source                                       |
| <i>BEL</i>        | Converts the original pathway files to BEL   |
| <i>Summarize</i>  | Presents global statistics of the total number of nodes and edges converted to BEL         |
| <i>Statistics</i> | Creates an excel sheet that summarizes the results of the BEL conversion for every pathway |



**Fig. 2** **a** PathMe Viewer main page. **b** The merged *mTOR* signaling network from KEGG, Reactome, and WikiPathways visualized in the PathMe Viewer. The highlighted regions mark the boundaries of each of the networks to definitively identify pathway landscapes, as defined by each of the network sources

host documentation, pytest (<https://github.com/pytest-dev/pytest>) as a unit and integration testing harness, and Travis-CI as a continuous integration server to run each of these with each commit (<https://travis-ci.com/PathwayMerger/PathMe> and <https://travis-ci.com/PathwayMerger/PathMe-Viewer>). Each package is distributed publicly through PyPI such that they can be included in other Python projects with requirements.txt or included in the setup.py using the *install\_requires* setting without the need for complicated build steps or any other user configuration.

Because PathMe works on frequently updated external pathway data from multiple sources, it must be re-run frequently to incorporate those updates. Following the recommendation from Kim et al. [28] for building reproducible environments for bioinformatics, we have encapsulated the entire PathMe workflow of acquiring, parsing, mapping, and normalizing the pathway resources within a Docker container such that it can be run on a cron job (i.e. a task scheduled to be re-run periodically). After, these changes are incorporated into the publicly available instance of the PathMe-Viewer. The cron job has the additional benefit that it reports when the formats of the underlying data change (which happens with moderate frequency) so the relevant PathMe components can be adapted. Also following the recommendation from Kim et al. for the scientific aspect of reproducibility, the three application scenarios presented in the next sections were conducted in IPython notebooks that are available and documented on GitHub (<https://github.com/PathwayMerger/PathMe-Resources>) that illustrate useful commands that might serve to assist similar future analyses.

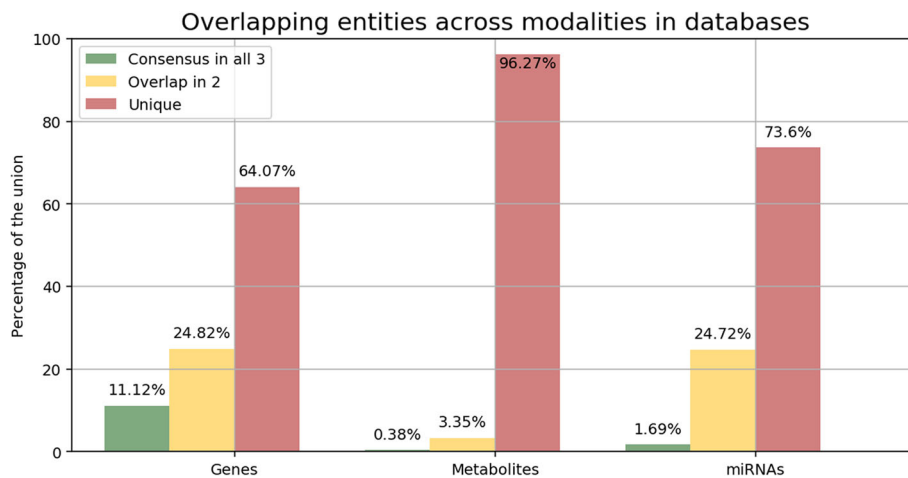
### Case scenario I: global entity comparison across pathway databases

As a first application, we conducted a global comparison of biological entities across major modalities (Fig. 3). We would like to note that in order to ensure the quality of the comparison presented in this case scenario, this analysis exclusively uses a highly cited and peer-reviewed pathway set provided by WikiPathways (approximately 510) that has been approved and tagged for usability in data analysis. While we attempted to maximize the retention of biological entities, we found severe differences in the level of overlap across resources which demonstrates the importance of database integration to gain a holistic picture of pathway knowledge.

The degree of consensus of biological entities across all three databases was found to be relatively low, albeit variable across the assessed modalities (Fig. 3). The proportion of genes present in all databases was lower than the results obtained by Stobbe et al. (15%), though they exclusively focused their work on a set of metabolic pathways present in five major databases which included KEGG and Reactome. Total consensus of miRNAs in all three databases was unsurprisingly low due to a disproportionate representation of miRNA species across the databases. Specifically, as few as 13 miRNAs were derived from Reactome while 149 were present in KEGG. Similarly, the total consensus for metabolites was grossly deficient at less than 2%.

For partial overlap, we found that results varied across the three modalities, with a higher degree of overlap between miRNA species at approximately 30%, followed by genes with nearly 20%, and metabolites with approximately 11%. Accordingly, we found the proportion of





**Fig. 3** Overlapping entities across modalities in the three databases studied (i.e., KEGG, Reactome, and WikiPathways). The comparison analysis studied the degree of overlap for three different biological entities (i.e., genes, metabolites, and miRNAs) to evaluate whether entities are shared across databases (i.e., the ratio of the number of nodes present in all three databases to the number of nodes in the union of all databases for that modality), partially overlap (i.e., the ratio of the number of nodes present in only two databases to the number of nodes in the union of all databases for that modality) or are exclusive (i.e., the ratio of the number of nodes unique in one database to the number of nodes in the union of all databases for that modality). The classification of entities by their corresponding modalities are described in Table 2. We would like to remark that the analysis accounted for every entity present in the full set of pathways from the studied databases

distinct entities to be substantially higher than those present in any two or all three databases. The particularly low levels of overlap observed in all modalities can be largely attributed to several factors:

- 1. The number of entities per modality across databases is highly variable** (Table 2). Per definition, sets with significant variations in cardinality (i.e., set size) limit the likelihood of consensus since only a portion of the larger sets can intersect with the smaller ones. For instance, KEGG contains 4048 metabolites while WikiPathways only contains 655. Accordingly, the maximum number of metabolites that can be common among them is limited to the number of metabolites contained in WikiPathways (i.e., 655). In this case, the maximum overlap would be the total number of metabolites contained in WikiPathways divided by the total number contained in KEGG, or 16.18%. Thus, the maximum degree of consensus between databases can be constrained by databases which contain fewer entities.
- 2. The scope of the pathways comprised in each database varies.** Each database places a distinct emphasis on discrete aspects or regions of biological pathways which tend to be defined subjectively in the absence of standard nomenclatures, as outlined by [14] who reported only 21 equivalent pathways between the three

databases. Therefore, despite the presence of key biological players in all three databases, the majority of biological entities are particular to a single database. For example, over 200 glycan molecules are present in KEGG since this resource contains multiple pathways related to glycan metabolism (i.e., 'Glycan biosynthesis and metabolism') while they are absent in the others.

- 3. Highly specific entity identifiers impede entity mappings with major standard nomenclatures.** Some entity identifiers have no discernible mapping

**Table 2** Pathway database content statistics. Each cell reflects the unique number of entities for a given modality in its corresponding database. The *genes* modality comprises genes, mRNAs, and gene products as well as any modifications on those. The *metabolites* modality comprises biological entities from small molecules to cellular components. The *miRNAs* modality contains microRNA molecules. Finally, nodes that correspond to other pathways, molecular events, or biological processes (e.g., Gene Ontology (GO; [8]) terms) are included in the *biological processes* modality. The statistics reflect the status of the content available from KEGG and WikiPathways from the 13th of March, 2019 and the latest Reactome release (version 67) from the 13th of December, 2018

| Modality             | KEGG | Reactome | WikiPathways |
|----------------------|------|----------|--------------|
| Genes                | 7289 | 8653     | 3361         |
| Metabolites          | 4048 | 2712     | 655          |
| miRNAs               | 149  | 13       | 91           |
| Biological processes | 418  | 2219     | 138          |

to major standard nomenclatures because they exhibit a high degree of specificity. This is particularly evident for genes where curators have also captured entities such as specific protein events (e.g., “p-y641-stat6”), protein mutations (e.g., “activated fgfr1 mutants”), protein families (e.g., “lim kinases”), protein fragments (e.g., “ub c-terminal hh fragments”), or splicing events (e.g., “xbp1 mrna (spliced)”). A way to account for these high granular cases would be to standardize protein family names with resources such as Pfam [16] or FamPlex [3]. For cases such as protein fragments or events, BEL enables their harmonization as they can be incorporated into its syntax (e.g., BEL *proteinModification()*, *fragment()*, *variant()*, etc.).

#### 4. Biological modalities can be broadly defined.

We characterize modalities to correspond to BEL node classes (Table 2). For instance, the *genes* modality comprises gene, protein, and RNA BEL nodes. While this modality is clearly defined, there is a higher degree of variability in the entity types that can be that can be classified with the *metabolites* modality since the latter comprises a broad range of abundance BEL nodes (i.e., small molecules, cellular components, clinical measurements, or categories that do not fit in other BEL node classes; [34]). Without the use of standard nomenclatures by the source databases, an extensive manual effort would be required to partition these modalities into more granular classifications. For example, the usage of GO as opposed to internally-defined terminologies to define cellular components would enable the categorization of cellular components into their own distinct modality. Similarly, the *biological processes* modality exhibits minimal overlap due to a lack of usage of standardized ontologies such as GO (Additional file 1).

#### Case scenario II: comparing equivalent pathways in the three databases

Merging pathway knowledge enables analyzing the crosstalks for any set of pathways through the PathMe Viewer. As a case scenario, we used PathMe in conjunction with the viewer to explore the knowledge consolidated from 21 equivalent pathways across the three databases previously curated by Domingo-Fernández et al. (Table 3). While conducting a cross-database pathway comparison previously required either extensive manual curation or harmonization of both entity identifiers and data formats on a case by case basis, this example illustrates how PathMe can be

exploited to enable a systematic comparison of equivalent pathways.

To evaluate the degree of overlap between the three representations of each equivalent pathway, we used a variation of the Szymkiewicz–Simpson coefficient calculated for the common molecular nodes between the networks (Eq. 1).

Each of the 21 equivalent pathways showed partial overlap, except ‘Non-homologous end-joining’ which did not contain the pathway information required to convert the pathway into BEL in two of its original files. Among the equivalent pathways with the highest degree of similarity, we found well-studied pathways such as ‘Cell cycle’, ‘Toll-like receptor signaling’, ‘mTOR signaling’, ‘Hedgehog signaling’, and ‘Apoptosis’. Although the three databases represent the most widely studied molecular players in each of these pathways, merging their knowledge assists in filling the gaps between the complex interactions occurring in these pathways. Pathways with low similarity, such as ‘TCA Cycle’ and ‘Sphingolipid Metabolism’, indicate the resources captured distinct aspects of the biology within the pathway. Unsurprisingly, this is in concordance with the findings reported by Stobbe et al. who conducted a comparison of the ‘TCA Cycle’ across five metabolic pathway databases. We would like to note that while previous approaches to characterize pathway similarity were purely gene-centric, our approach includes not only gene sets, but a range of modalities represented in pathways. Finally, beyond harmonizing entities and concepts, PathMe also harmonizes relationships, thus facilitating further analyses where pathway topology is included, as shown in the next case scenario.

#### Case scenario III: in-depth pathway analysis of mTOR signaling after superimposing its multiple representations

As a further application of the framework, we used the PathMe Viewer to conduct a detailed investigation of the mammalian target of rapamycin (mTOR) signaling pathway to demonstrate its utility in enriching pathway knowledge. In Fig. 4, the consensus in terms of entity overlap across equivalent mTOR signaling pathways from each of the databases is depicted. All three databases are complementary to the others, but also possess some degree of overlap and thus are neither entirely identical nor distinct. Variability in the size of the mTOR signaling pathway, as measured by the number of nodes in each database, is also clearly discernible with KEGG contributing the largest proportion of distinct nodes to the heterogeneous, merged network (Fig. 4a).

A key functionality of PathMe Viewer is in the visualization and interactive exploration of pathways. In Fig. 5a and b, using the viewer, an in-depth analysis of mTOR signaling reveals novel sets of interactions in

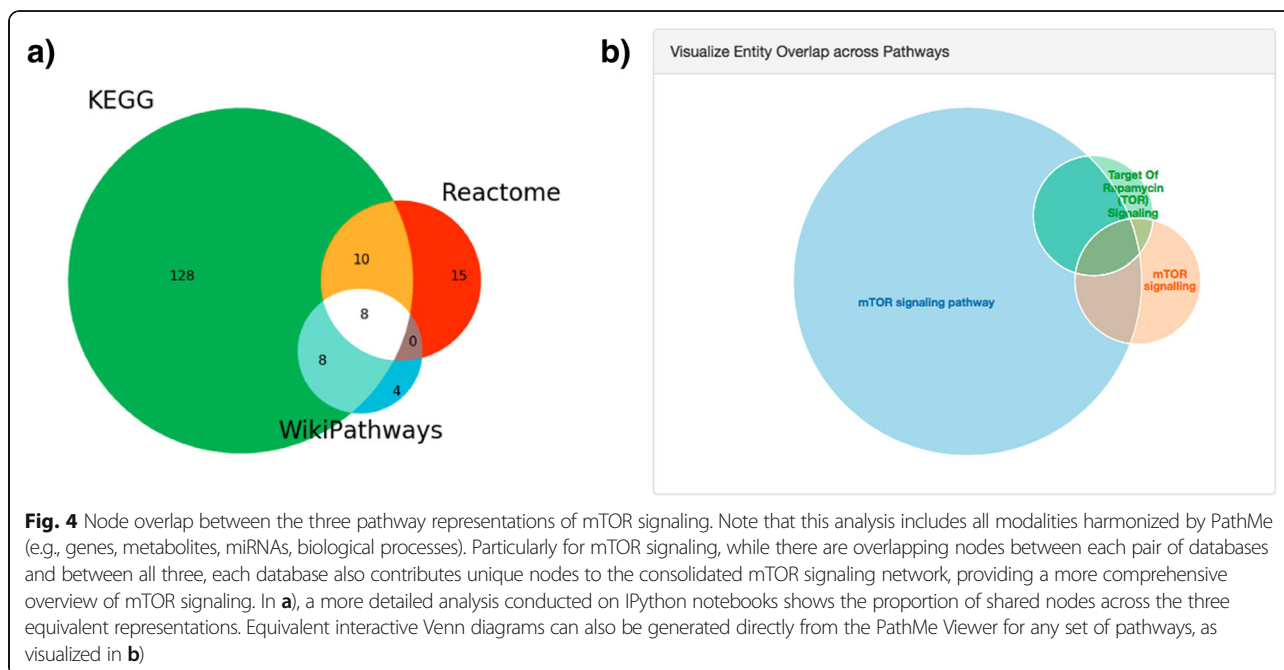
**Table 3** Consolidated pathway representations, their similarity indexes, and links to visualize the merged networks in the PathMe Viewer. A detailed analysis with the scripts to replicate the results and comments on the identified overlaps for each of the 21 equivalent pathways is available at [https://nbviewer.jupyter.org/github/PathwayMerger/PathMe-Resources/blob/master/notebooks/case\\_scenarios/evaluating\\_similarity\\_equivalent\\_pathways.ipynb](https://nbviewer.jupyter.org/github/PathwayMerger/PathMe-Resources/blob/master/notebooks/case_scenarios/evaluating_similarity_equivalent_pathways.ipynb)

| KEGG                                       | Reactome   | WikiPathways                               | Pathway Similarity Index |
|--|--|--|--------------------------|
| Cell cycle                                 | Cell Cycle   | Cell Cycle                                 | 0.70                     |
| Toll-like receptor signaling pathway       | Toll-Like Receptors Cascades                           | Toll-like Receptor Signaling Pathway       | 0.62                     |
| mTOR signaling pathway                     | mTOR signalling  | Target Of Rapamycin (TOR) Signaling        | 0.58                     |
| Hedgehog signaling pathway                 | Signaling by Hedgehog                                  | Hedgehog Signaling Pathway                 | 0.56                     |
| Apoptosis                                  | Apoptosis  | Apoptosis                                  | 0.43                     |
| IL-17 signaling pathway                    | Interleukin-17 signaling                               | IL17 signaling pathway                     | 0.42                     |
| PI3K-Akt signaling pathway                 | PI3K/AKT activation                                    | PI3K-Akt Signaling Pathway                 | 0.42                     |
| Wnt signaling pathway                      | Signaling by WNT                                       | Wnt Signaling Pathway                      | 0.41                     |
| MAPK signaling pathway                     | MAPK family signaling cascades                         | MAPK Signaling Pathway                     | 0.40                     |
| B cell receptor signaling pathway          | B Cell Receptor Signaling Pathway                      | Signaling by the B Cell Receptor (BCR)     | 0.37                     |
| Pentose phosphate pathway                  | Pentose phosphate pathway (hexose monophosphate shunt) | Pentose Phosphate Pathway                  | 0.33                     |
| Citrate cycle (TCA cycle)                  | Citric acid cycle (TCA cycle)                          | TCA Cycle                                  | 0.33                     |
| Synthesis and degradation of ketone bodies | Ketone body metabolism                                 | Synthesis and Degradation of Ketone Bodies | 0.33                     |
| Notch signaling pathway                    | Signaling by NOTCH                                     | Notch Signaling Pathway                    | 0.29                     |
| DNA replication                            | DNA Replication  | DNA Replication                            | 0.28                     |
| Prolactin signaling pathway                | Prolactin receptor signaling                           | Prolactin receptor signaling               | 0.28                     |
| TGF-beta signaling pathway                 | Signaling by TGF-beta family members                   | TGF-beta Signaling Pathway                 | 0.26                     |
| Thyroid hormone synthesis                  | Thyroxine biosynthesis                                 | Thyroxine (Thyroid Hormone) Production     | 0.20                     |
| Sphingolipid metabolism                    | Sphingolipid metabolism                                | Sphingolipid Metabolism                    | 0.16                     |
| Mismatch repair                            | Mismatch Repair  | Mismatch repair                            | 0.08                     |
| Non-homologous end-joining                 | Nonhomologous End-Joining (NHEJ)                       | Non-homologous end joining                 | 0                        |

the integrated network that are absent in individual mTOR signaling networks. The role of AKT signaling in modulating mTOR, as illustrated in 5a and sourced from KEGG, has already been well-described in the literature [1, 33]. More notably, by superimposing the mTOR signaling pathway as defined in KEGG with its equivalent pathway from WikiPathways (Fig. 5b), an association between AKT1 and insulin related-processes becomes apparent in the merged network, though neither of the individual pathway sources connect the downstream effects of mTOR on insulin signaling. Nevertheless, the association between mTOR and insulin signaling through AKT modulation has been previously described in the literature [30]. Additionally, bidirectional effects of mTOR have been demonstrated on AKT activity; these effects can vary both by the type of mTOR complex involved in the pathway and by negative feedback loops on insulin signaling, leading to altered states of AKT activity [1, 30]. As such, both pathways serve to complement each other in the integrated network, unraveling a connection which was

hidden in disparate databases, though has been well-studied in the literature. Recent studies have also demonstrated that mTOR receives input from multiple pathways [33]; a principal feature of the PathMe Viewer is in its capacity to directly visualize pathway crosstalk. While in the previous case scenario, crosstalk analyses were performed across equivalent pathways, in this case, using the viewer it would be possible to simultaneously visualize and explore different pathways which are evidenced to, or are possibly involved in, crosstalk with the mTOR signaling one.

Similarly, by superimposing the mTOR signaling pathways from KEGG and WikiPathways, downstream interactions between mTOR and mRNA translation via EIF4EBP1 are evident (Fig. 5a and b). The inhibition of mTOR has been noted to be a potent repressor of protein translation while mTOR activation can stimulate mRNA translation through EIF4EBP1 [10, 19]. Though only the inhibition relationship is captured in the viewer, in the presence of an activation relationship, the viewer also offers a feature to detect contradictory

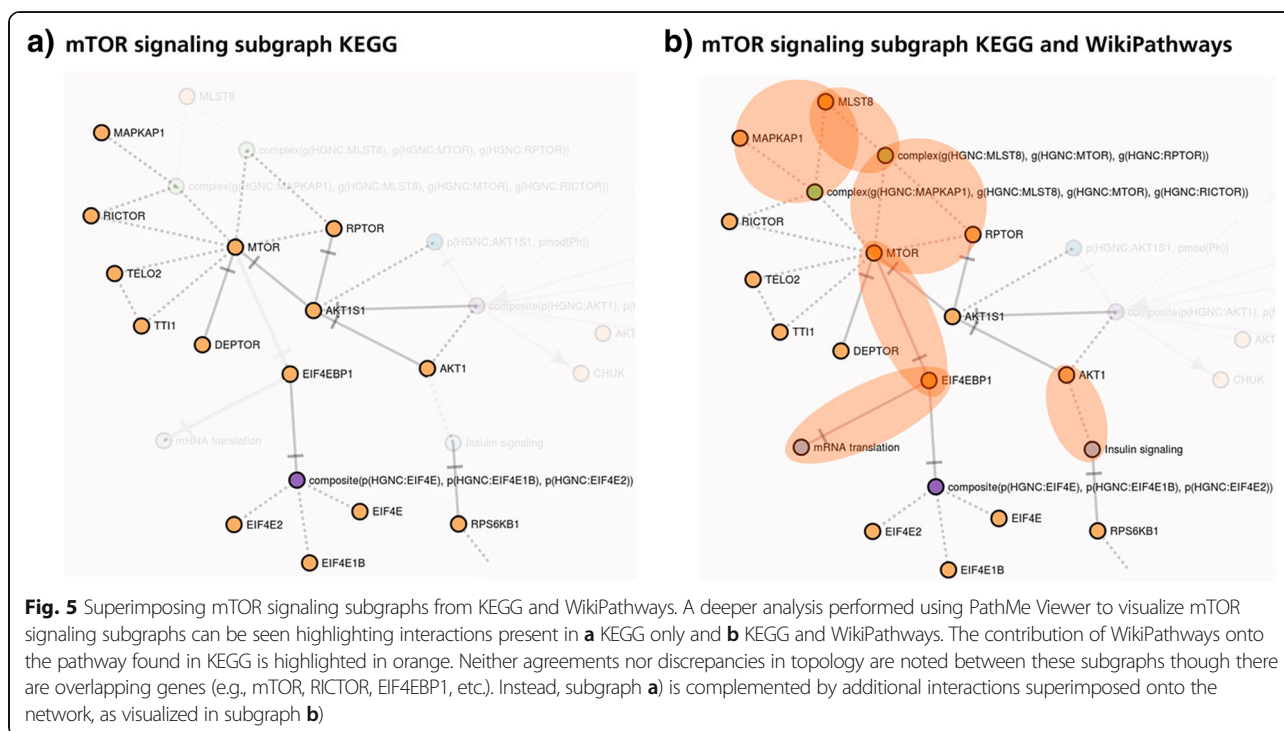


edges (e.g., node A increases node B in one pathway and decreases B in another) between identical nodes across two databases.

**Conclusions**

Parallel developments of pathway databases during recent decades have resulted in different formalization schemas, hampering the interoperability between these

resources and creating data silos. Overcoming this obstacle is instrumental to better understand the mechanisms underlying pathway knowledge. Additionally, while our approach can accommodate multi-scale pathway information from divergent database formats into a singular and standardized schema, a minority of entities and interactions have no discernible equivalencies in BEL and, as such, had to be omitted. For instance, so far



PathMe parsers can extract information from both humans and other species; however, despite the capacity of PathMe to harmonize human identifiers, additional work is required for the harmonization of identifiers belonging to other species as integration can help in identifying evolutionarily conserved genes and processes.

Here, we have presented a framework through which content across multiple pathway databases can be integrated and transformed into a unified schema. Although PathMe currently only incorporates content from three major pathway databases, its flexibility allows for future inclusion of additional pathway databases. Moreover, it holds the capacity to update its content and track developments in pathway knowledge, an issue earlier outlined by Wadi et al.. Finally, the three case scenarios presented illustrate how the framework can be used to assist researchers in addressing biological questions at varying degrees of specificity such as: i) integrating the pathway landscape at the database level, ii) comparing the degree of consensus at the pathway level, and iii) exploring pathway crosstalk and studying consensus at the molecular level.

Ultimately, we have shown how integrating pathway databases and making them interoperable enables global pathway representations that can contribute to a more holistic overview of pathway knowledge than the knowledge contained in any single one of the databases. In the future, these global representations could be used to conduct more comprehensive pathway-centric analyses. Furthermore, the reproducibility of previous pathway enrichment analyses could also be evaluated by replicating them using any database combination. In other words, what would happen if, instead of KEGG, an identical analysis were to be performed using the Reactome or WikiPathways databases, or any combination of the three?

## Availability and requirements

**Project name:** PathMe

**Project home page:** <https://github.com/PathwayMerger>

**Operating system(s):** Platform independent

**Programming language:** Python and JavaScript

**Other Requirements:** Python 3

**License:** Apache License 2.0

**Any restrictions to use by non-academics:** KEGG Commercial License

## Additional file

**Additional file 1:** Supplementary Text. (PDF 442 kb)

## Abbreviations

BEL: Biological Expression Language; BioPAX: Biological Pathway Exchange; DSL: Domain Specific Language; GO: Gene Ontology; MVC: Model-View-Controller; PyPI: Python Package Index; RDF: Resource Description Framework; SBGN: Systems Biology Graphical Notation; SBML: Systems

Biology Markup Language; SPIA: Signaling Pathway Impact Analysis; XML: Extensible Markup Language

## Acknowledgements

We are very grateful to the curators of KEGG, Reactome, and WikiPathways for generating the raw content which was used in this work. Furthermore, we would like to thank Dr. Egon Willighagen for his helpful suggestions regarding WikiPathways data.

## Funding

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

The funding body did not play a role in the design of the study and collection, analysis, and interpretation of data, or in writing the manuscript.

## Availability of data and materials

The datasets generated and/or analysed during the current study are available in the ComPath's GitHub repository, [<https://github.com/ComPath/resources>]. The datasets generated and/or analysed during the current study are publicly available at <https://github.com/PathwayMerger/PathMe-Resources>.

## Authors' contributions

DDF conceived and designed the study. SM and JML implemented the individual database parsers with help and supervision from DDF. CTH supervised harmonization into BEL using the PyBEL framework. DDF implemented the web application and conducted the application scenario with the help of SM. DDF, SM, and CTH wrote the paper. MHA participated in the critical definition of the concept, proposed and participated in drafting and commenting critically the manuscript. All authors have read and approved the final manuscript.

## Ethics approval and consent to participate

Not applicable.

## Consent for publication

Not applicable.

## Competing interests

The authors declare that they have no competing interests.

## Publisher's Note

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

Received: 5 March 2019 Accepted: 29 April 2019

Published online: 15 May 2019

## References

- Altomare AD, Khaled RA. Homeostasis and the importance for a balance between AKT/mTOR activity and intracellular signaling. *Curr Med Chem.* 2012;19(22):3748–62. <https://doi.org/10.2174/092986712801661130>.
- Apweiler R, et al. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 2004;32(suppl\_1):D115–9. <https://doi.org/10.1093/nar/gkh131>.
- Bachman JA, Gyori BM, Sorger PK. FamPlex: a resource for entity recognition and relationship resolution of human protein families and complexes in biomedical text mining. *BMC Bioinformatics.* 2018;19(1):248. <https://doi.org/10.1186/s12859-018-2211-5>.
- Belinky F, et al. PathCards: multi-source consolidation of human biological pathways. *Database.* 2015;2015. <https://doi.org/10.1093/database/bav006>.
- Bohler A, et al. Reactome from a WikiPathways perspective. *PLoS Comput Biol.* 2016;12(5):e1004941. <https://doi.org/10.1371/journal.pcbi.1004941>.
- Bolton EE, Wang Y, Thiessen PA, Bryant SH. PubChem: integrated platform of small molecules and biological activities. *Ann Reports Computational Chem.* 2008;4:217–41. [https://doi.org/10.1016/S1574-1400\(08\)00012-1](https://doi.org/10.1016/S1574-1400(08)00012-1).
- Bonnet E, et al. BiNoM 2.0, a Cytoscape plugin for accessing and analyzing pathways using standard systems biology formats. *BMC Syst Biol.* 2013;7(1):18. <https://doi.org/10.1186/1752-0509-7-18>.

8. Carbon S, et al. Expansion of the gene ontology knowledgebase and resources: the gene ontology consortium. *Nucleic Acids Res.* 2017;45(D1):331–8 <https://doi.org/10.1093/nar/gkw1108>.
9. Cerami EG, et al. Pathway commons, a web resource for biological pathway data. *Nucleic Acids Res.* 2011;39(Suppl. 1):D685–90 <https://doi.org/10.1093/nar/gkq1039>.
10. Choo AY, Yoon SO, Kim SG, Roux PP, Blenis J. Rapamycin differentially inhibits S6Ks and 4E-BP1 to mediate cell-type-specific repression of mRNA translation. *Proc Natl Acad Sci.* 2008;105(45):17414–9 <https://doi.org/10.1073/pnas.0809136105>.
11. Chowdhury S, Sarkar RR. Comparison of human cell signaling pathway databases—evolution, drawbacks and challenges. *Database.* 2015.
12. Demir E, et al. The BioPAX community standard for pathway data sharing. *Nat Biotechnol.* 2010;28(9):935 <https://doi.org/10.1038/nbt.1666>.
13. Demir E, et al. Using biological pathway data with paxtools. *PLoS Comput Biol.* 2013;9(9):e1003194 <https://doi.org/10.1371/journal.pcbi.1003194>.
14. Domingo-Fernández D, et al. ComPath: an ecosystem for exploring, analyzing, and curating mappings across pathway databases. *NPJ Syst Biol Appl.* 2018;4(1): 43 <https://doi.org/10.1038/s41540-018-0078-8>.
15. Fabregat A, et al. The Reactome pathway knowledgebase. *Nucleic Acids Res.* 2018;46(D1):D649–55 <https://doi.org/10.1093/nar/gkx1132>.
16. Finn RD, et al. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* 2015;44(D1):D279–85 <https://doi.org/10.1093/nar/gkv1344>.
17. Franz M, et al. Cytoscape.js: a graph theory library for visualisation and analysis. *Bioinformatics.* 2015;32(2):309–11 <https://doi.org/10.1093/bioinformatics/btv557>.
18. Gaulton A, et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* 2011;40(D1):D1100–7 <https://doi.org/10.1093/nar/gkr777>.
19. Gingras AC, et al. Hierarchical phosphorylation of the translation inhibitor 4E-BP1. *Genes Dev.* 2001;15(21):2852–64 <https://doi.org/10.1101/gad.912401>.
20. Gyori BM, et al. From word models to executable models of signaling networks using automated assembly. *Mol Syst Biol.* 2017;13(11):954 <https://doi.org/10.15252/msb.20177651>.
21. Hastings J, et al. ChEBI in 2016: improved services and an expanding collection of metabolites. *Nucleic Acids Res.* 2015;44(D1):D1214–9 <https://doi.org/10.1093/nar/gkv1031>.
22. Hoyt CT, Konotopez A, Ebeling C. PyBEL: a computational framework for biological expression language. *Bioinformatics.* 2017;34(4):703–4 <https://doi.org/10.1093/bioinformatics/btx660>.
23. Hoyt CT, et al. BEL commons: an environment for exploration and analysis of networks encoded in biological expression language. *Database.* 2018; 2018:bay126 <https://doi.org/10.1093/database/bay126>.
24. Hubbard T, et al. The Ensembl genome database project. *Nucleic Acids Res.* 2002;30(1):38–41.
25. Hucka M, et al. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics.* 2003;19(4):524–31 <https://doi.org/10.1093/bioinformatics/btg015>.
26. Kamburov A, Wierling C, Lehrach H, Herwig R. ConsensusPathDB—a database for integrating human functional interaction networks. *Nucleic Acids Res.* 2008; 37(suppl\_1):D623–8 <https://doi.org/10.1093/nar/gkn698>.
27. Kanehisa M, et al. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 2016;45(D1):D353–61 <https://doi.org/10.1093/nar/gkw1092>.
28. Kim YM, Poline JB, Dumas G. Experimenting with reproducibility: a case study of robustness in bioinformatics. *GigaScience.* 2018;7(7):gij077 <https://doi.org/10.1093/gigascience/gij077>.
29. Kutmon M, et al. PathVisio 3: An Extendable Pathway Analysis Toolbox. *PLoS Comput Biol.* 2015;11(2):e1004085 <https://doi.org/10.1371/journal.pcbi.1004085>.
30. Le Bacquer O, et al. mTORC1 and mTORC2 regulate insulin secretion through Akt in INS-1 cells. *J Endocrinol.* 2013;216(1):21–9 <https://doi.org/10.1530/JOE-12-0351>.
31. Le Novère N, et al. The systems biology graphical notation. *Nat Biotechnol.* 2009;27(8):735 <https://doi.org/10.1038/nbt.1558>.
32. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez gene: gene-centered information at NCBI. *Nucleic Acids Res.* 2005;33(suppl\_1):D54–8 <https://doi.org/10.1093/nar/gki031>.
33. Memmott RM, Dennis PA. Akt-dependent and-independent mechanisms of mTOR regulation in cancer. *Cell Signal.* 2009;21(5):656–64 <https://doi.org/10.1016/j.cellsig.2009.01.004>.
34. Pham N, et al. Consistency, inconsistency, and ambiguity of metabolite names in biochemical databases used for genome-scale metabolic modelling. *Metabolites.* 2019;9:28 <https://doi.org/10.3390/metabo9020028>.
35. Povey S, et al. The HUGO gene nomenclature committee (HGNC). *Hum Genet.* 2001;109(6):678–80 <https://doi.org/10.1007/s00439-001-0615-0>.
36. Pratt D, Chen J, Welker D, et al. NDEX, the network data exchange. *Cell Systems.* 2015;1(4):302–5 <https://doi.org/10.1016/j.cels.2015.10.001>.
37. Sales G, et al. metaGraphite - a new layer of pathway annotation to get metabolite networks. *Bioinformatics.* 2018;bty719 <https://doi.org/10.1093/bioinformatics/bty719>.
38. Slenker N, et al. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res.* 2017;46(D1): D661–7 <https://doi.org/10.1093/nar/gkx1064>.
39. Stobbe MD, et al. Critical assessment of human metabolic pathway databases: a stepping stone for future integration. *BMC Syst Biol.* 2011;5(1): 165 <https://doi.org/10.1186/1752-0509-5-165>.
40. Tarca AL, et al. A novel signaling pathway impact analysis. *Bioinformatics.* 2008;25(1):75–82 <https://doi.org/10.1093/bioinformatics/btn577>.
41. Türei D, Korcsmáros T, Saez-Rodriguez J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat Methods.* 2016;13(12): 966 <https://doi.org/10.1038/nmeth.4077>.
42. Iersel v, et al. The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. *BMC Bioinformatics.* 2010;11(1):5 <https://doi.org/10.1186/1471-2105-11-5>.
43. Wadi L, et al. Impact of outdated gene annotations on pathway enrichment analysis. *Nat Methods.* 2016;13(9):705 <https://doi.org/10.1038/nmeth.3963>.
44. Wrzodek C, Dräger A, Zell A. KEGGtranslator: visualizing and converting the KEGG PATHWAY database to various formats. *Bioinformatics.* 2011;27(16): 2314–5 <https://doi.org/10.1093/bioinformatics/btr377>.
45. Zhang B, et al. RaMP: a comprehensive relational database of metabolomics pathways for pathway enrichment analysis of genes and metabolites. *Metabolites.* 2018;8(1):16 <https://doi.org/10.3390/metabo8010016>.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://www.biomedcentral.com/submissions)



## Conclusions

We have presented PathMe [120], the first tool that successfully harmonizes pathway networks from major databases. We also implemented PathMe Viewer [121], a complementary web application that provides comprehensive network visualization of the consensus pathway knowledge. This entire framework adheres to the same principles of flexibility and reproducibility as ComPath (chapter 2).

In the paper, we demonstrate that consolidating pathway knowledge from multiple resources results in more comprehensive pathway representations. By overlaying equivalent pathways across databases with the help of the mappings curated with ComPath, we were able to derive their consensus networks. From there, we uncovered novel relationships and contradictory evidences within equivalent pathways. Furthermore, using PathMe, we carried out a systematic evaluation of the similarity across the pathway landscape. Previously, this entire process would have been done manually [61].

In the future, additional databases can be incorporated into the framework thanks to its flexible design. Since the results of pathway-based analyses are heavily influenced by the dynamic changes in pathway knowledge, we have also implemented PathMe to automatically update its content [122]. Finally, this work leaves two open questions. First, if the results of a pathway-driven analysis are influenced by pathway choice, and second, whether the results improve by using integrative resources versus individual databases.





# 4 The impact of pathway database choice on statistical enrichment analysis and predictive modeling

## Introduction

The abundance of pathway databases has resulted in an unintended segmentation of knowledge. Researchers can only gain close familiarity with a small number of these databases. As a result, they limit their attention to a popular few, and the knowledge contained in the many other databases remains out of sight. In fact, despite this plethora of resources, most pathway analyses are still conducted using a single database. The choice of this database is the result of the researcher's own biases, preferences, and experience.

However, as illustrated in the last two chapters, not only is the overlap of pathways across databases low, but even representations of the same biological pathway can notably differ. Building on the success of the tools discussed in the two previous chapters, we present the first benchmarking study evaluating the choice of pathway database on a broad spectrum of pathway enrichment methods and predictive modeling applications. This study demonstrates the significant influence of database selection in the downstream results of enrichment methods as well as in the performance of predictive models.

Reprinted with permission from "Sarah Mubeen, Charles Tapley Hoyt, André Gemünd, Martin Hofmann-Apitius, Holger Fröhlich, and Daniel Domingo-Fernández. The impact of pathway database choice on statistical enrichment analysis and predictive modeling. *Frontiers in Genetics*, 10:1203". Copyright © Daniel Domingo-Fernández 2019.



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

Sarah Mubeen<sup>1,2</sup>, Charles Tapley Hoyt<sup>1,2†</sup>, André Gemünd<sup>1</sup>, Martin Hofmann-Apitius<sup>1,2</sup>, Holger Fröhlich<sup>2</sup> and Daniel Domingo-Fernández<sup>1,2\*</sup>

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

## OPEN ACCESS

### Edited by:

Lavanya Balakrishnan,  
Mazumdar Shaw Medical Centre,  
India

### Reviewed by:

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

### \*Correspondence:

Daniel Domingo-Fernández  
danieldomingofernandez@hotmail.com

### †ORCID:

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

### Specialty section:

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

**Received:** 23 August 2019

**Accepted:** 30 October 2019

**Published:** 22 November 2019

### Citation:

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

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

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

## INTRODUCTION

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

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

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

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

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

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

## MATERIALS AND METHODS

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

### Pathway Databases

#### Selection Criteria

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

#### Data Retrieval and Processing

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

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

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

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

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

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

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



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

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

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

## Clinical and Genomic Data

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

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

## Pathway Enrichment Methods

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

### Overrepresentation Analysis

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

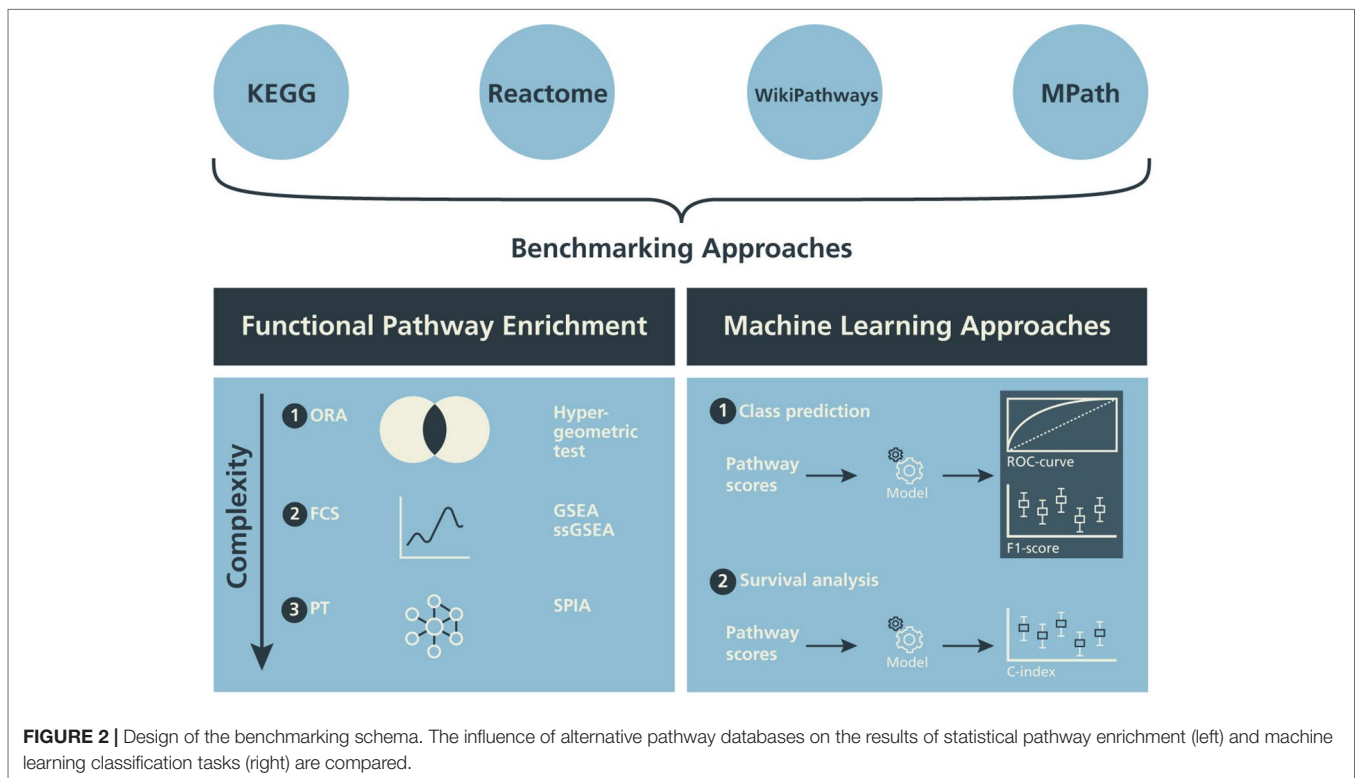
### Functional Class Scoring Methods

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

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

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

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



### Pathway Topology-Based Enrichment

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

### Evaluation Based on Enrichment of Pathway Analogs

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

### Machine Learning

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

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

#### Prediction of Tumor vs. Normal

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

#### Prediction of Tumor Subtype

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

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

### Prediction of Overall Survival

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

### Statistical Assessment of Database Impact on Prediction Performance

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

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

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

### Software Implementation

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

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

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

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

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

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

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

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



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

## Hardware

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

## RESULTS

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

## Benchmarking the Impact on Enrichment Methods

### Overrepresentation Analysis

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

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

### Gene Set Enrichment Analysis

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

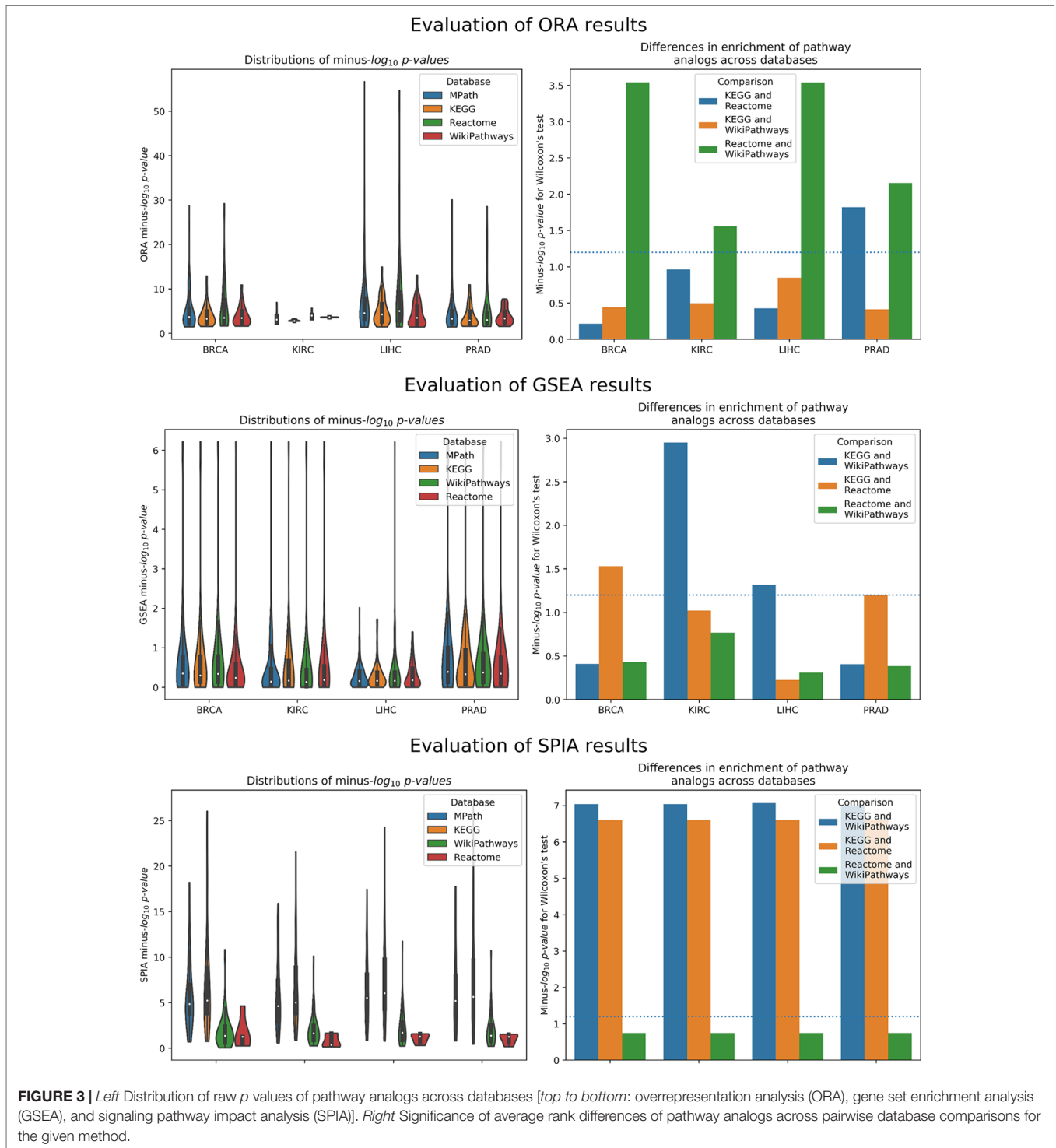
### Signaling Pathway Impact Analysis

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

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

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

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

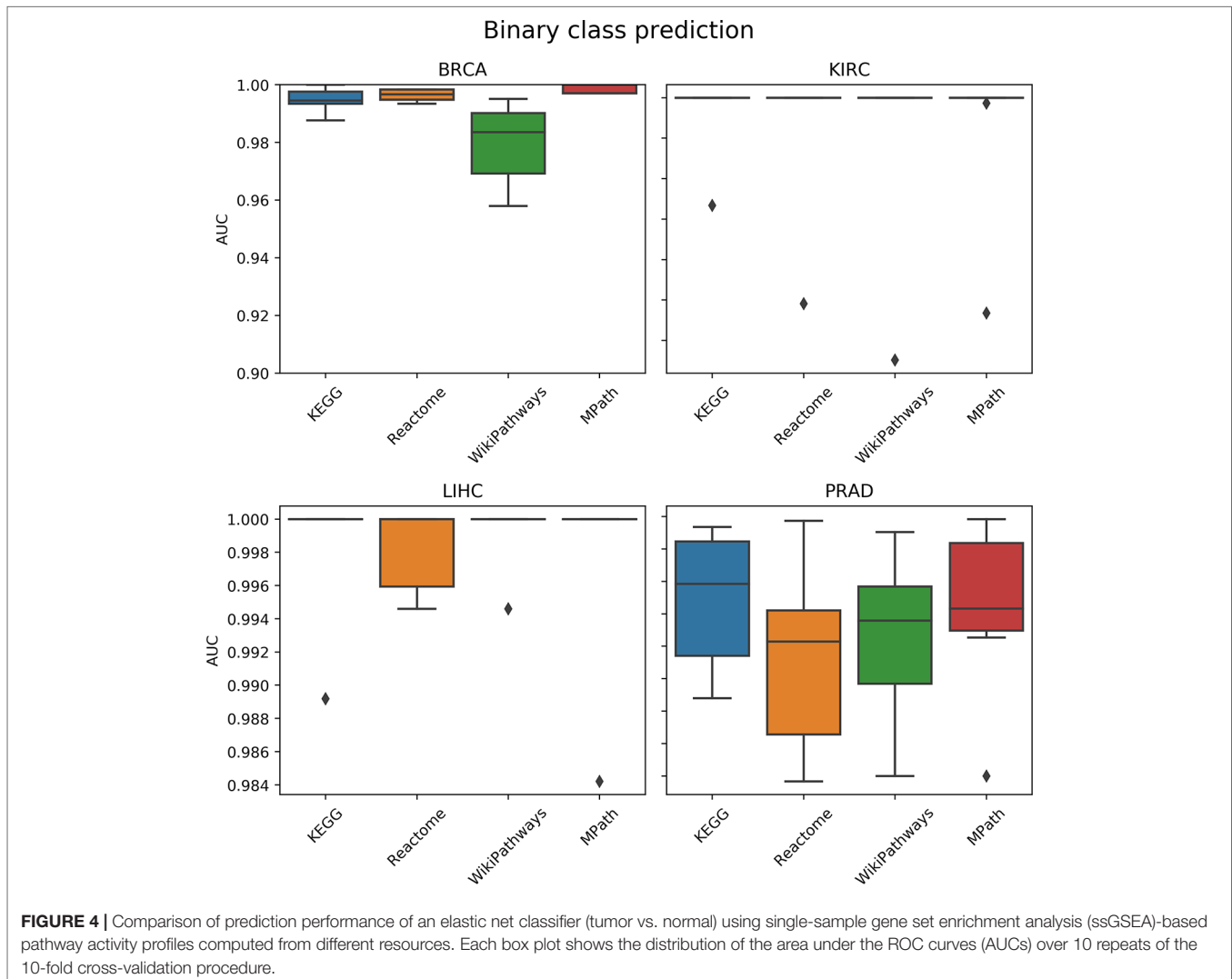


was activated and the WikiPathways one inhibited). A similar effect was observed in the "Estrogen signaling pathway," found to be inhibited in KEGG and activated in WikiPathways in the LIHC dataset. The merging of equivalent pathway networks resulted in the observation of inhibition for both the "TGF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

## Benchmarking the Impact on Predictive Modeling

### Prediction of Tumor vs. Normal

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



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

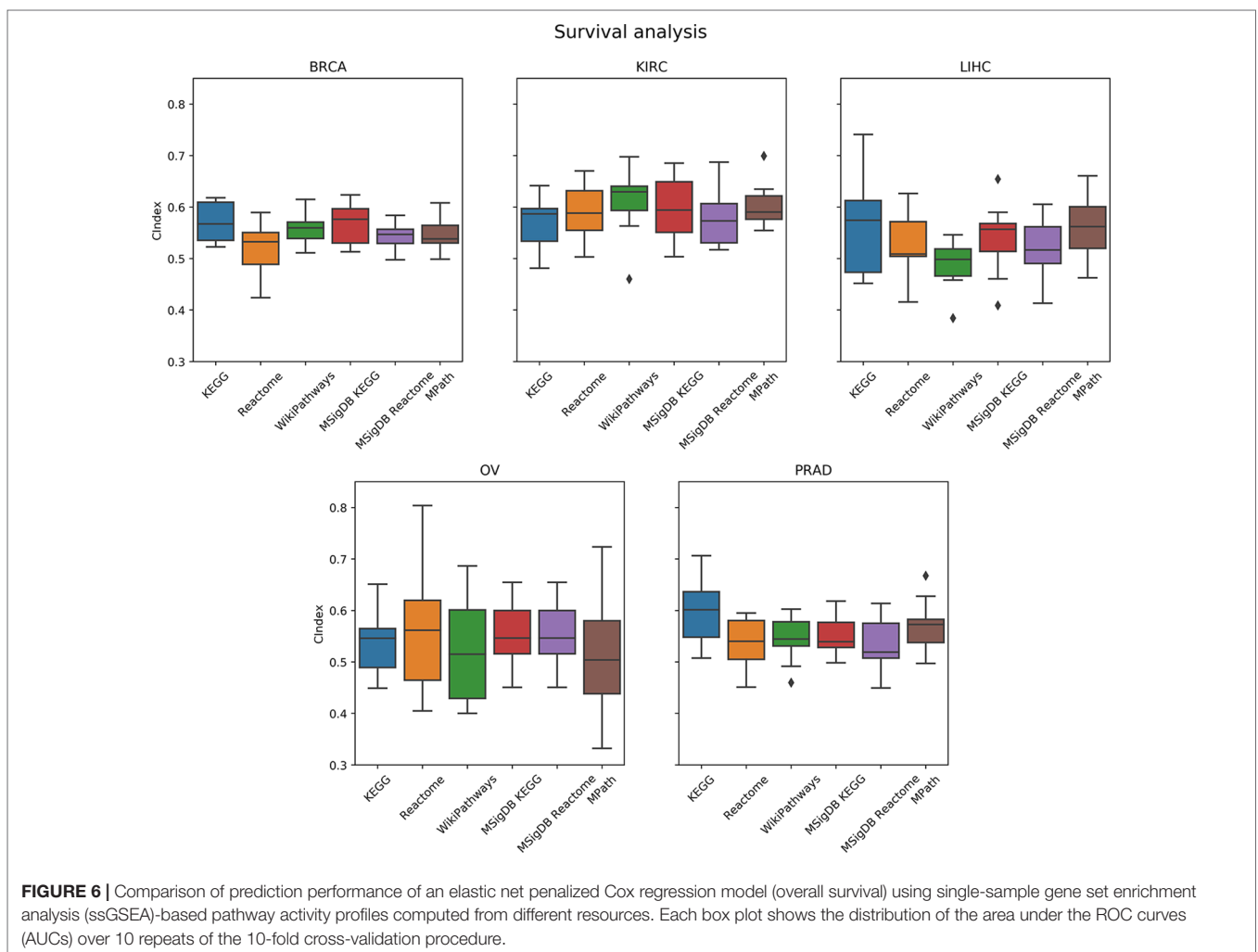
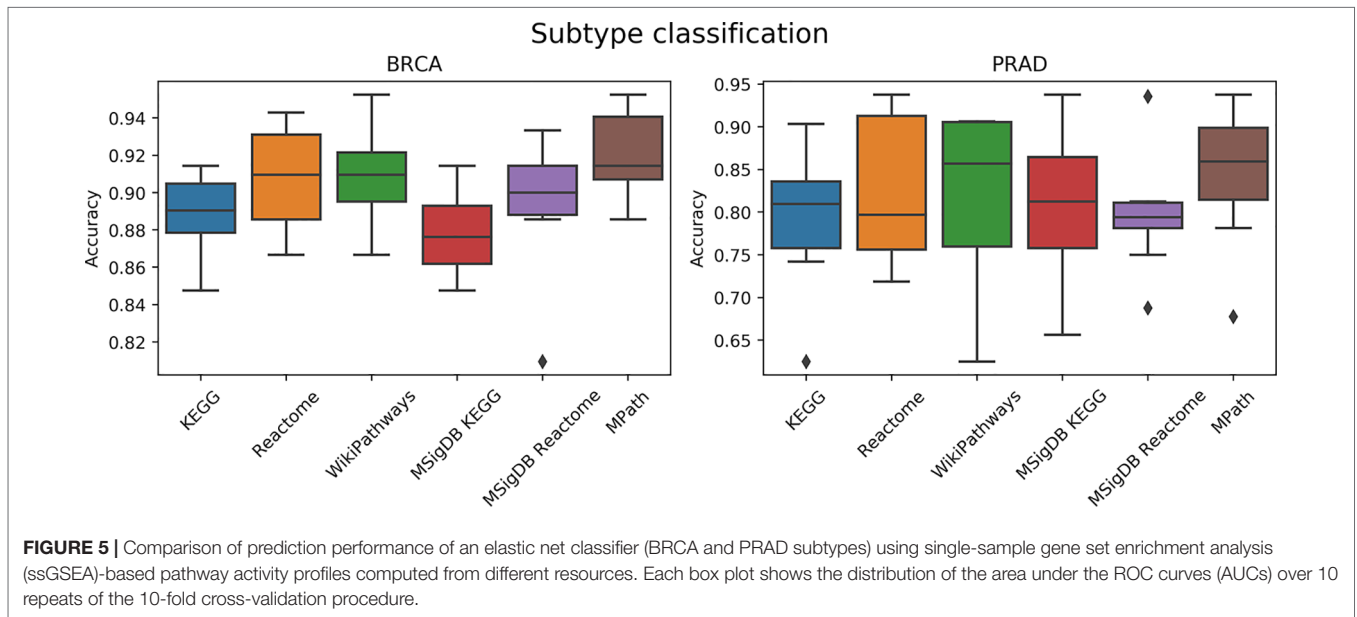
### Prediction of Tumor Subtype

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

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

### Prediction of Overall Survival

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



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

## DISCUSSION

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

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

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

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

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

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

## DATA AVAILABILITY STATEMENT

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

## AUTHOR CONTRIBUTIONS

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

## FUNDING

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

## ACKNOWLEDGMENTS

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

## SUPPLEMENTARY MATERIAL

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

## REFERENCES

- Bader, G. D., Cary, M. P., and Sander, C. (2006). Pathguide: a pathway resource list. *Nucleic Acids Res.* 34 (suppl\_1), D504–D506. doi: 10.1093/nar/gkj126
- Barbie, D. A., Tamayo, P., Boehm, J. S., Kim, S. Y., Moody, S. E., Dunn, I. F., et al. (2009). Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* 462 (7269), 108. doi: 10.1038/nature08460
- Bateman, A. R., El-Hachem, N., Beck, A. H., Aerts, H. J., and Haibe-Kains, B. (2014). Importance of collection in gene set enrichment analysis of drug response in cancer cell lines. *Sci. Rep.* 4, 4092. doi: 10.1038/srep04092
- Bayerlová, M., Jung, K., Kramer, F., Klemm, F., Bleckmann, A., and Beißbarth, T. (2015). Comparative study on gene set and pathway topology-based enrichment methods. *BMC Bioinf.* 16 (1), 334. doi: 10.1186/s12859-015-0751-5
- Belinky, F., Nativ, N., Stelzer, G., Zimmerman, S., Iny Stein, T., Safran, M., et al. (2015). PathCards: multi-source consolidation of human biological pathways. *Database* 2015. doi: 10.1093/database/bav006
- Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B. (Methodological)* 57 (1), 289–300. doi: 10.2307/2346101
- Benjamini, Y., and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* 29 (4), 1165–1188. doi: 10.1214/aos/1013699998
- Cancer Genome Atlas Research Network. (2014). Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* 513 (7517), 202. doi: 10.1038/nature13480
- Cerami, E. G., Gross, B. E., Demir, E., Rodchenkov, I., Babur, O., Anwar, N., et al. (2011). Pathway commons, a web resource for biological pathway data. *Nucleic Acids Res.* 39 (Suppl. 1), D685–D690. doi: 10.1093/nar/gkq1039
- Coates, A. S., Winer, E. P., Goldhirsch, A., Gelber, R. D., Gnant, M., Piccart-Gebhart, M., et al. (2015). Tailoring therapies—improving the management of early breast cancer: St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2015. *Ann. Oncol.* 26 (8), 1533–1546. doi: 10.1093/annonc/mdv221
- Colaprico, A., Silva, T. C., Olsen, C., Garofano, L., Cava, C., Garolini, D., et al. (2015). TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* 44 (8), e71–e71. doi: 10.1093/nar/gkv1507
- Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* 20 (3), 273–297. doi: 10.1007%2FBF00994018
- Doderer, M. S., Anguiano, Z., Suresh, U., Dashnamoorthy, R., Bishop, A. J., and Chen, Y. (2012). Pathway Distiller-multisource biological pathway consolidation. *BMC Genom.* 13 (6), S18. doi: 10.1186/1471-2164-13-S6-S18
- Domingo-Fernández, D., Hoyt, C. T., Bobis-Álvarez, C., Marin-Llao, J., and Hofmann-Apitius, M. (2018). ComPath: an ecosystem for exploring, analyzing, and curating mappings across pathway databases. *NPJ Syst. Biol. Appl.* 4 (1), 43. doi: 10.1038/s41540-018-0078-8
- Domingo-Fernandez, D., Mubeen, S., Marin-Llao, J., Hoyt, C., and Hofmann-Apitius, M. (2019). PathMe: merging and exploring mechanistic pathway knowledge. *BMC Bioinf.* 20, 243. doi: 10.1186/s12859-019-2863-9
- Drier, Y., Sheffer, M., and Domany, E. (2013). Pathway-based personalized analysis of cancer. *Proc. Nat. Acad. Sci.* 110 (16), 6388–6393. doi: 10.1073/pnas.1219651110
- Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., et al. (2018). The reactome pathway knowledgebase. *Nucleic Acids Res.* 46 (D1), D649–D655. doi: 10.1093/nar/gkx1132
- Fabris, F., Palmer, D., de Magalhães, J. P., and Freitas, A. A. (2019). Comparing enrichment analysis and machine learning for identifying gene properties that discriminate between gene classes. *Briefings Bioinf.* doi: 10.1093/bib/bbz028
- Fisher, R. A. (1992). Statistical methods for research workers in *Breakthroughs in Statistics* (New York, NY:Springer), 66–70.
- Fröhlich, H. (2014). Including network knowledge into Cox regression models for biomarker signature discovery. *Biom. J.* 56 (2), 287–306. doi: 10.1002/bimj.201300035
- Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Software* 33 (1), 1. doi: 10.18637/jss.v033.i01
- García-Campos, M. A., Espinal-Enríquez, J., and Hernández-Lemus, E. (2015). Pathway analysis: state of the art. *Front. Physiol.* 6, 383. doi: 10.3389/fphys.2015.00383
- Grüning, B. A., Lampa, S., Vaudel, M., and Blankenberg, D. (2019). Software engineering for scientific big data analysis. *GigaScience* 8 (5), giz054. doi: 10.1093/gigascience/giz054
- Graudenzi, A., et al. (2017). Pathway-based classification of breast cancer subtypes. *Front. Biosci., (Landmark Ed)* 22, 1697–1712. doi: 10.2741/4566
- Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., and Rosati, R. A. (1982). Evaluating the yield of medical tests. *JAMA* 247 (18), 2543–2546. doi: 10.1001/jama.1982.03320430047030
- Hoyt, C. T., Konotopez, A., and Ebeling, C. (2018). PyBEL: a computational framework for Biological Expression Language. *Bioinformatics* 34 (4), 703–704. doi: 10.1093/bioinformatics/btx660
- Hoyt, C. T., Domingo-Fernández, D., Mubeen, S., Llaó, J. M., Konotopez, A., Ebeling, C., et al. (2019). Integration of Structured Biological Data Sources using Biological Expression Language. *Biorxiv* 631812. doi: 10.1101/631812
- Ihnatova, I., Popovici, V., and Budinska, E. (2018). A critical comparison of topology-based pathway analysis methods. *PLoS One* 13 (1), e0191154. doi: 10.1371/journal.pone.0191154
- Kamburov, A., Wierling, C., Lehrach, H., and Herwig, R. (2008). ConsensusPathDB—a database for integrating human functional interaction networks. *Nucleic Acids Res.* 37 (suppl\_1), D623–D628. doi: 10.1093/nar/gkn698
- Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2016). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 45 (D1), D353–D361. doi: 10.1093/nar/gkw1092
- Khatri, P., Sirota, M., and Butte, A. J. (2012). Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.* 8 (2), e1002375. doi: 10.1371/journal.pcbi.1002375
- Kirouac, D. C., Saez-Rodriguez, J., Swantek, J., Burke, J. M., Lauffenburger, D. A., and Sorger, P. K. (2012). Creating and analyzing pathway and protein interaction compendia for modelling signal transduction networks. *BMC Syst. Biol.* 6 (1), 29. doi: 10.1186/1752-0509-6-29
- Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. (2015). The molecular signatures database hallmark gene set collection. *Cell Syst.* 1 (6), 417–425. doi: 10.1016/j.cels.2015.12.004
- Lim, S., Lee, S., Jung, I., Rhee, S., and Kim, S. (2018). Comprehensive and critical evaluation of individualized pathway activity measurement tools on pan-cancer data. *Briefings Bioinf.*
- Lim, S., Park, Y., Hur, B., Kim, M., Han, W., and Kim, S. (2016). Protein interaction network (pin)-based breast cancer subsystem identification and activation measurement for prognostic modeling. *Methods* 110, 81–89. doi: 10.1016/j.ymeth.2016.06.015
- Mayr, A., and Schmid, M. (2014). Boosting the concordance index for survival data—a unified framework to derive and evaluate biomarker combinations. *PLoS One* 9 (1), e84483. doi: 10.1371/journal.pone.0084483
- McKinney, W. (2010). Data Structures for Statistical Computing in Python in *Proceedings of the 9th Python in Science Conference*. Eds. van der Walt, S., and Millman, J., 51–56.
- Miller, R. A., Ehrhart, F., Eijssen, L. M., Slenter, D. N., Curfs, L. M., Evelo, C. T., et al. (2019). Beyond pathway analysis: Identification of active subnetworks in Rett syndrome. *Front. Genet.* 10, 59. doi: 10.3389/fgene.2019.00059
- Molinari, A. M., Simon, R., and Pfeiffer, R. M. (2005). Prediction error estimation: a comparison of resampling methods. *Bioinformatics* 21 (15), 3301–3307. doi: 10.1093/bioinformatics/bti499
- Povey, S., Lovering, R., Bruford, E., Wright, M., Lush, M., and Wain, H. (2001). The HUGO gene nomenclature committee (HGNC). *Hum. Genet.* 109 (6), 678–680. doi: 10.1007/s00439-001-0615-0
- Reimand, J., Isserlin, R., Voisin, V., Kucera, M., Tannus-Lopes, C., Rostamianfar, A., et al. (2019). Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nat. Protoc.* 14 (2), 482–517. doi: 10.1038/s41596-018-0103-9
- Sales, G., Calura, E., and Romualdi, C. (2018). meta Graphite—a new layer of pathway annotation to get metabolite networks. *Bioinformatics* 35 (7), 1258–1260. doi: 10.1093/bioinformatics/bty719

- Schaefer, C. F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., et al. (2008). PID: the pathway interaction database. *Nucleic Acids Res.* 37 (suppl\_1), D674–D679. doi: 10.1093/nar/gkn653
- Senkus, E., Kyriakides, S., Ohno, S., Penault-Llorca, F., Poortmans, P., Rutgers, E., et al. (2015). Primary breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.* 26 (suppl\_5), v8–v30. doi: 10.1093/annonc/mdv298
- Slater, T. (2014). Recent advances in modeling languages for pathway maps and computable biological networks. *Drug Discovery Today* 19 (2), 193–198. doi: 10.1016/j.drudis.2013.12.011
- Slenter, D. N., Kutmon, M., Hanspers, K., Riutta, A., Windsor, J., Nunes, N., et al. (2017). WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res.* 46 (D1), D661–D667. doi: 10.1093/nar/gkx1064
- Sorlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci.* 98 (19), 10869–10874. doi: 10.1073/pnas.191367098
- Stoney, R. A., Schwartz, J. M., Robertson, D. L., and Nenadic, G. (2018). Using set theory to reduce redundancy in pathway sets. *BMC Bioinf.* 19 (1), 386. doi: 10.1186/s12859-018-2355-3
- Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Nat. Acad. Sci.* 102 (43), 15545–15550. doi: 10.1073/pnas.0506580102
- Tarca, A. L., Draghici, S., Khatri, P., Hassan, S. S., Mittal, P., Kim, J. S., et al. (2008). A novel signaling pathway impact analysis. *Bioinformatics* 25 (1), 75–82. doi: 10.1093/bioinformatics/btn577
- Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Stat. Med.* 16 (4), 385–395. doi: 10.1002/(sici)1097-0258(19970228)16:4<385::aid-sim380>3.0.co;2-3
- Türei, D., Korcsmáros, T., and Saez-Rodriguez, J. (2016). OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat. Methods* 13 (12), 966. doi:10.1038/nmeth.4077
- Van Wieringen, W. N., Kun, D., Hampel, R., and Boulesteix, A. L. (2009). Survival prediction using gene expression data: a review and comparison. *Comput. Stat. Data Anal.* 53 (5), 1590–1603. doi: 10.1016/j.csda.2008.05.021
- Vivar, J. C., Pemu, P., McPherson, R., and Ghosh, S. (2013). Redundancy control in pathway databases (ReCiPa): an application for improving gene-set enrichment analysis in Omics studies and "Big data" biology. *Omics: J. Integr. Biol.* 17 (8), 414–422. doi: 10.1089/omi.2012.0083
- Wadi, L., Meyer, M., Weiser, J., Stein, L. D., and Reimand, J. (2016). Impact of outdated gene annotations on pathway enrichment analysis. *Nat. Methods* 13 (9), 705. doi: 10.1038/nmeth.3963
- Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* 45 (10), 1113. doi: 10.1038/ng.2764
- Zhang, Y., Yang, W., Li, D., Yang, J. Y., Guan, R., and Yang, M. Q. (2018). Toward the precision breast cancer survival prediction utilizing combined whole genome-wide expression and somatic mutation analysis. *BMC Med. Genom.* 11 (5), 104. doi: 10.1109/BIBM.2017.8217762
- Zou, H., and Trevor, H. (2005). Regularization and Variable Selection via the Elastic Net. *J. R. Stat. Soc. Ser. B: 67* (2), 301–320. doi: 10.1111/j.1467-9868.2005.00503.x

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

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

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





## Conclusions

This work is the first benchmarking study that highlights the significant impact of database selection on statistical pathway enrichment and predictive modeling. There are three major messages to be drawn from this work. First, we have confirmed that results of pathway-driven approaches can be heavily influenced by database selection. These findings suggest that future pathway-driven analyses should be further validated with a complementary database (e.g., replicating the results with a second database). Second, it has also demonstrated that the observed differences at database level can be mitigated by using integrative approaches such as MPath. We have demonstrated how our unifying approach outperforms the predictive power of models that use individual databases. This point is highly relevant to the machine and deep learning field, where a vast amount of data is required. Third, we have shown a possible and practical solution (i.e., MPath's construction workflow) for future integration of pathway resources by leveraging the work presented in the previous chapters (i.e., ComPath and PathMe). While centralizing approaches are essential to break down data silos, they are also fundamental to drive research by exploiting the explanatory power derived from pathway knowledge.



# 5 Multimodal Mechanistic Signatures for Neurodegenerative Diseases (NeuroMMSig): a web server for mechanism enrichment

## Introduction

After decades of research, many clinical trials, and billions invested, there is still no treatment for the major neurological disorders such as AD or PD. This suggests that it could be time to take a step back and analyze the potential mistakes made, before launching yet another study (doomed to fail?). In the AD area, for instance, the pharmaceutical industry still tries to target two of the main mechanisms implicated in AD (i.e., amyloid beta and tau protein), despite decades of continuous failures [87]. Instead of focusing on single mechanisms, we could try to understand how all the numerous mechanisms that are in the aetiology of these disorders could lead to the disease state. Following this direction, we present Multimodal Mechanistic Signatures for Neurodegenerative Diseases (NeuroMMSig), the largest inventory of knowledge-based mechanisms for AD and PD. In NeuroMMSig, each of the mechanisms is formalized as a computable network and exposed through a web application that enables the interpretation of multi-scale and multimodal clinical data by adopting the novel paradigm of mechanism enrichment.

Reprinted with permission from "Daniel Domingo-Fernández *et al.*. Multi-modal Mechanistic Signatures for Neurodegenerative Diseases (NeuroMMSig): a web server for mechanism enrichment. *Bioinformatics*, Volume 33, Issue 22, 15 November 2017, Pages 3679–3681". Copyright © Daniel Domingo-Fernández 2017.

Databases and ontologies

# Multimodal mechanistic signatures for neurodegenerative diseases (NeuroMMSig): a web server for mechanism enrichment

Daniel Domingo-Fernández<sup>1,2</sup>, Alpha Tom Kodamullil<sup>1,2</sup>,  
Anandhi Iyappan<sup>1,2</sup>, Mufassra Naz<sup>1,2</sup>, Mohammad Asif Emon<sup>1,2</sup>,  
Tamara Raschka<sup>1,2</sup>, Reagon Karki<sup>1,2</sup>, Stephan Springstube<sup>1</sup>,  
Christian Ebeling<sup>1</sup> and Martin Hofmann-Apitius<sup>1,2,\*</sup>

<sup>1</sup>Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing, Sankt Augustin 53754, Germany and <sup>2</sup>Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn-Aachen International Center for IT, Bonn 53113, Germany

\*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on April 6, 2017; revised on May 24, 2017; editorial decision on June 12, 2017; accepted on June 21, 2017

## Abstract

**Motivation:** The concept of a ‘mechanism-based taxonomy of human disease’ is currently replacing the outdated paradigm of diseases classified by clinical appearance. We have tackled the paradigm of mechanism-based patient subgroup identification in the challenging area of research on neurodegenerative diseases.

**Results:** We have developed a knowledge base representing essential pathophysiology mechanisms of neurodegenerative diseases. Together with dedicated algorithms, this knowledge base forms the basis for a ‘mechanism-enrichment server’ that supports the mechanistic interpretation of multiscale, multimodal clinical data.

**Availability and implementation:** NeuroMMSig is available at <http://neurommsig.scai.fraunhofer.de/>

**Contact:** martin.hofmann-apitius@scai.fraunhofer.de

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

## 1 Introduction

The development of novel high throughput ‘omic’ technologies in the last decade has revealed new insight and progresses in areas of cancer, cardiovascular and metabolic disorders. The datasets coming from these technologies have led to the discovery of candidate biomarkers and potential drug targets. However, in other areas such as neurodegenerative diseases, this mechanistic understanding is either rather limited or almost absent.

Readouts in translational biomedicine are going beyond molecular level: they can span from genes and genetic variation information to imaging and organ-level (or even organism-level) data and markers. The definition of a disease as ‘dysregulated pathways’ may

hold true for cancer, but is inappropriate for neurodegenerative diseases as pathways refer typically to rapid molecular processes and the alterations in neurodegenerative diseases are slow and multifaceted. There is simply no such thing as a ‘degeno-gene’ (in analogy to the ‘onco-gene’). Supporting that, there have not been any described ‘cause-effect’ relationships that would explain the different pathological changes observed in patients with these disorders. When the effects of dysregulation can be easily observed—like in monogenic diseases—it is generally not so difficult to link the phenotype with the event that lead to it. This is likely to be attributed to a short and direct chain of causality (Hofmann-Apitius *et al.*, 2015a). Hence, because the complexity of neurodegenerative diseases is

enormous; it is crucial to integrate a wider spectrum of causal assertions into models that represent and organize the available mechanistic knowledge.

MSigDB (Subramanian *et al.*, 2015) is the prototypic implementation of a system that allows for the identification of perturbed pathways. However, the output of ranking algorithms like GSEA is usually a list of associated canonical pathways that do not contain disease-specific information and multimodal data. In addition, canonical pathways are also biased towards cancer biology (Hofmann-Apitius *et al.*, 2015b).

Adopting the fundamental principle of ‘running patterns in data against a knowledge base of established patterns’ (‘pathways’; ‘signatures’), we have developed a mechanism enrichment server and extended it towards multiscale and multimodal data. This is where the two ‘M’ of NeuroMMSig come from: Multimodal and Mechanistic. It is noteworthy that the difference between NeuroMMSig and other, conventional methods for pathway enrichment or functional gene annotation lies in the specificity of the disease context. Pathway enrichment is based upon canonical pathways, which are not disease specific. The multimodal mechanisms behind NeuroMMSig, however, are manually curated and contain detailed representations of multimodal pathophysiology in a well-defined disease context.

Here, we present NeuroMMSig, a web server for mechanism enrichment that allows submission of multiscale data from molecular to clinical level to return mechanisms that fit best the data. We have focused on neurodegenerative diseases, as we try to establish a ‘mechanism-based taxonomy of Alzheimer’s Disease (AD) and Parkinson’s Disease (PD)’. This is the core of the AETIONOMY project ([www.aetionomy.eu](http://www.aetionomy.eu)) and in fact, NeuroMMSig (DB and Server) form the backbone of attempts at stratifying patient subgroups based on disease mechanisms.

## 2 Systems and methods

### 2.1 Categorization of NDD pathways from mechanism based models

Disease knowledge assembly models were built using Biological Expression Language (BEL) which integrate literature-derived ‘cause and effect’ relationships in the form of triples (Kodamullil *et al.*, 2015). We have captured a representative subsample of the scientific knowledge on existing canonical pathways in AD and PD (Iyappan *et al.*, 2016) which have been grouped into subgraphs.

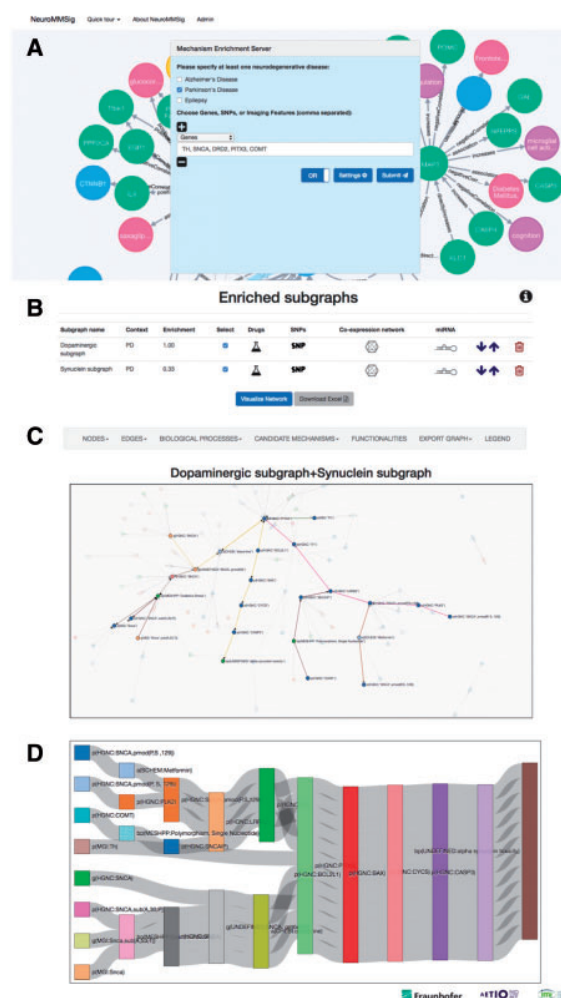
### 2.2 Multimodal data integration, data sources and software

NeuroMMSig’s subgraphs have been enriched with multimodal data (e.g. imaging features, variant information and drugs). The methodology describing how the linking across different data scales was performed is provided in the Supplementary text. Moreover, we have developed an enrichment algorithm to rank the subgraphs based on the input.

## 3 Implementation

### 3.1 NeuroMMSig server

NeuroMMSig is available at <http://neurommsig.scai.fraunhofer.de/>. A user interface offers a simple, yet comprehensive menu (Fig. 1A). Input fields allow users to submit multimodal data (e.g. genes, SNPs, imaging features). Users can also set the enrichment algorithm parameters and define the operators of the query. After data



**Fig. 1.** The web interface for NeuroMMSig. (A) Home page. Users can select the disease context and submit their data. In the navbar, links are provided to ‘Introduction’ and ‘How to use’ pages. (B) Example of the results provided by ranking algorithm with the most enriched subgraphs. (C) By clicking ‘Visualize Network’, the users can access the interactive network visualization with multiple functionalities. (D) Sankey diagram of causal relationships comprising the proposed mechanism

submission, a ranked list of subgraphs is displayed to the user (Fig. 1B). Here, associated information to the submitted data is shown as icons in a user friendly table: drug-gene interactions, known regulating miRNA and co-expressed networks. Moreover, when the user selects one or multiple subgraphs and clicks on ‘Visualize Network’, NeuroMMSig displays the graph representing the selection where the user can investigate how the disruption of the network occurs (Fig. 1C). For that reason, NeuroMMSig offers multiple functionalities enabling graph mining and reasoning over the graphs (e.g. graph algorithms, search and exporting options, knowledge provenance and Sankey diagram representations for pathway analysis).

### 3.2 Application scenario

The five most relevant genes associated with ‘Dopamine signaling pathway’ in PD according to SCAIView [<http://academia.scaiview.com/academia/>] (Supplementary text) were used as an input (Fig. 1A). Two subgraphs were retrieved from NeuroMMSig

(‘Dopaminergic subgraph’ and ‘Synuclein subgraph’) and they were selected for further analysis (Fig. 1B). Using the query tools, the two main hub nodes SNCA and Parkinson’s disease were removed from the network in order to avoid most of the paths going through them, which biases the retrieval of best candidate mechanisms. By choosing a process of interest such as ‘alpha synuclein toxicity’, the server proposes candidate mechanisms in which the data-mapped-nodes may perturb normal physiology (Fig. 1C and D).

#### 4 Discussion

Harmonization of heterogeneous and multiscale datasets is yet a tremendous challenge in the field of neurodegeneration. The gap between molecular and clinical data is too wide to establish stable and meaningful assertions between imaging features and genes, for instance. Thus, integration of different data scales is a necessary step to shed some light on the mechanisms underlying neurodegenerative diseases.

The modeling approach chosen in NeuroMMSig is capable of explaining causal and correlative relationships among different entities namely genes, proteins or biological processes in the context of neurological disorders (Kodamullil *et al.*, 2015). These relationships reveal the upstream and downstream regulators of each node in the network and how they are activating/inhibiting their neighboring nodes. Thus, navigating through the network it is possible to identify the root or primarily cause of a dysfunctional gene or protein which eventually contributes to the disorder.

The inventory of mechanisms specific for neurodegenerative diseases, which forms the basis of NeuroMMSig, is composed of small cause-and-effect models encoded in OpenBEL. Evidences for the BEL-encoded mechanisms come from the scientific literature, from experimental data analysis and from clinical readouts such as imaging biomarkers. Furthermore, both AD and PD models incorporate genetic and epigenetic information, which might, for instance, indicate and partially explain the effect of a particular SNP in a mechanism (Khanam *et al.*, 2015; Naz *et al.*, 2016). The presented work also serves as a comparison tool between different diseases. Thus, it allows to systematically identify shared-mechanisms between them. Combining all together, the BEL-encoded mechanisms contain pathophysiology information at highest resolution, with highly curated evidences spanning from the genetics and epigenetics layer via cell-type specific information to clinical phenotypes and biomarkers. Hence, NeuroMMSig overcomes some of challenges that pathway

analysis methods currently have, as indicated by Khatri *et al.* (2012).

#### Acknowledgements

We thank Andrej Konotopoz, Sumit Madan, André Gemünd and Charles Tapley Hoyt for technical assistance and valuable advices. We would also like to acknowledge Apurva Gopisetty and Anka Guldenpfennig for their support curating the models. Finally, we thank Shweta Bagewadi and Sepehr Golriz Khatami for their inputs to towards metadata inclusions.

#### Funding

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

*Conflict of Interest:* none declared.

#### References

- Hofmann-Apitius, M. *et al.* (2015a) Bioinformatics mining and modeling methods for the identification of disease mechanisms in neurodegenerative disorders. *Int. J. Mol. Sci.*, **16**, 29179–29206.
- Hofmann-Apitius, M. *et al.* (2015b) Towards the taxonomy of human disease. *Nat. Rev. Drug Discov.*, **14**, 75.
- Iyappan, A. *et al.* (2016) Towards a pathway inventory of the human brain for modeling disease mechanisms underlying neurodegeneration. *J. Alzheimer’s Dis.*, **52**, 1343–1360.
- Khanam, I.A. *et al.* (2015) Computational modelling approaches on epigenetic factors in neurodegenerative and autoimmune diseases and their mechanistic analysis. *J. Immunol. Res.*, **2015**.
- Khatri, P. *et al.* (2012) Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput Biol*, **8**, e1002375.
- Kodamullil, A.T. *et al.* (2015) Computable cause-and-effect models of healthy and Alzheimer’s disease states and their mechanistic differential analysis. *Alzheimer’s Dement.*, **11**, 1329–1339.
- Naz, M. *et al.* (2016) Reasoning over genetic variance information in cause-and-effect models of neurodegenerative diseases. *Brief. Bioinf.*, **17**, 505–516.
- Subramanian, A. *et al.* (2015) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.*, **102**, 15545–15550.





## Conclusions

This work presents a paradigm shift in the neurodegenerative field by offering to the scientific community over 200 networks that represent the knowledge around well-established disease-specific mechanisms involved in these disorders. The value of this resource has already been demonstrated in numerous applications spanning from drug discovery [123] to precision medicine (see chapter 7). One of the visions of precision medicine has been to re-define disease taxonomies based on molecular characteristics rather than on phenotypic evidence. This is precisely the most promising application of NeuroMMSig. Supported by its mechanistic backbone, we aim to build a mechanism-based taxonomy for AD and PD that can classify patient populations into different strata based on the mechanisms involved. Finally, the success and generalizability of this approach has prompted us to translate the concept of mechanism enrichment to other domains such as the psychiatric field.



# 6 PTSD Biomarker Database: deep dive meta-database for PTSD biomarkers, visualizations, and analysis tools

## Introduction

Although hundreds of biomarkers have been associated with PTSD, very few have been replicated, and none have yet been validated and qualified for deployment in clinical care. Additionally, biomarker-focused meta-analyses have systematically shown conflicting and inconsistent results [124–126]. Therefore, organizing and contextualizing published information and data is critical to understanding the level of confirmatory and contradictory evidence for single biomarkers. Moreover, there are no diagnostic biomarkers for this condition. Therefore, PTSD diagnosis is currently based solely on symptom presentation which often lead to late diagnosis. For these reasons, identifying diagnostic biomarkers is essential to an early intervention that can enable the evaluation and follow-up of disease progression. In this work, we present PTSD Biomarker Database (PTSDDB), the first resource aiming to capture and organize biomarker knowledge in this PTSD.

Reprinted with permission from "Daniel Domingo-Fernández *et al.*. PTSD Biomarker Database: Deep Dive Meta-database for PTSD Biomarkers, Visualizations, and Analysis Tools. *Database: The Journal of Biological Databases and Curation*, Volume 2019, baz081. Copyright © Daniel Domingo-Fernández 2019.



Original article

# PTSD Biomarker Database: deep dive metadatabase for PTSD biomarkers, visualizations and analysis tools

Daniel Domingo-Fernández<sup>1,†,\*</sup>, Allison Provost<sup>2,†</sup>,  
Alpha Tom Kodamullil<sup>1,†</sup>, Josep Marín-Llaó<sup>1</sup>, Heather Lasseter<sup>2</sup>,  
Kristophe Diaz<sup>2</sup>, Nikolaos P. Daskalakis<sup>2</sup>, Lee Lancashire<sup>2</sup>,  
Martin Hofmann-Apitius<sup>1</sup> and Magali Haas<sup>2</sup>

<sup>1</sup>Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing, Sankt Augustin 53754, Germany and <sup>2</sup>Cohen Veterans Bioscience, 1 Broadway, Cambridge, MA 02142, United States

\*Corresponding author: Tel: +49 2241 14-2354; Fax: +49 2241 14-2656; Email: daniel.domingo.fernandez@scai.fraunhofer.de

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

Citation details: Domingo-Fernández, D., Provost, A., Kodamullil, A. T. *et al.* PTSD Biomarker Database: deep dive metadatabase for PTSD biomarkers, visualizations and analysis tools. *Database* (2019) Vol. 2019: article ID baz081; doi:10.1093/database/baz081

Received 12 February 2019; Revised 12 May 2019; Accepted 28 May 2019

## Abstract

The PTSD Biomarker Database (PTSDDDB) is a database that provides a landscape view of physiological markers being studied as putative biomarkers in the current post-traumatic stress disorder (PTSD) literature to enable researchers to explore and compare findings quickly. The PTSDDDB currently contains over 900 biomarkers and their relevant information from 109 original articles published from 1997 to 2017. Further, the curated content stored in this database is complemented by a web application consisting of multiple interactive visualizations that enable the investigation of biomarker knowledge in PTSD (e.g. clinical study metadata, biomarker findings, experimental methods, etc.) by compiling results from biomarker studies to visualize the level of evidence for single biomarkers and across functional categories. This resource is the first attempt, to the best of our knowledge, to capture and organize biomarker and metadata in the area of PTSD for storage in a comprehensive database that may, in turn, facilitate future analysis and research in the field.

**Database URL:** <https://ptsd.scai.fraunhofer.de>

## Introduction

Post-traumatic stress disorder (PTSD) is a common psychiatric disorder that occurs in some individuals after a

traumatic event (13) and is diagnosed by mental health professionals based on the presentation of four symptom clusters—intrusions, avoidance, negative cognitions/mood and

hyperarousal (1). PTSD pathophysiology is complex and affects multiple interconnected biological systems that regulate mental and physical health functions and are associated with PTSD's clinical heterogeneity and diverse comorbidity profiles (2,5,9,11,14,15).

An extensive amount of research in PTSD has explored the utility of physiological markers as being discrete biomarkers of this disorder; however, no such putative biomarkers of PTSD have been identified to date based on the regulatory approval process for qualifying and validating biomarkers around specific clinical contexts of use (hereafter the term 'biomarker' will be used to refer to 'physiological markers of disease'). In the PTSD literature, the types of physiological markers most commonly studied include neuroimaging and psychophysiological measures, behavioral and neurocognitive readouts and analytes measured in peripheral biofluids, such as blood and saliva at baseline or after psychological challenge (4,6,8,12,20). Fluid-based peripheral biomarkers may include inflammation indicators, hypothalamic pituitary adrenal axis mediators, neurosteroids and neurotransmitters, which have functional roles both in the peripheral and central nervous system, potentially enabling biologically meaningful inference with clinical utility (4).

The increasing amount of biomarker studies in all disease areas is paralleled by the growing number of meta-analyses that combine data from multiple studies to systematically derive common conclusions. However, the lack of disease-specific biomarker registries impedes the harmonization and integration of results from these studies, which often remain in the form of non-structured text, figures, tables or supplementary files. Organizing and storing this knowledge is essential to provide a comprehensive view of the biomarker landscape and to foster the discovery and development of diagnostics and treatments.

Along these lines, several biomarker databases have been recently developed that focus on specific disease domains, such as colorectal cancer (19), Alzheimer's disease (10), tuberculosis (18) and liver cancer (3) to name a few. Furthermore, there are multiple resources that store and catalog biomarker information from multiple indications, such as the Online Mendelian Inheritance in Man (OMIM) (7), the cancer biomarker database (16) and the infectious disease database (17). These resources illustrate how biomarker information can be curated and harmonized and currently serve as hubs for biomarker research in their respective areas. Although biomarkers of PTSD—once identified—have the potential to improve patient outcomes, groups have yet to embark on similar efforts for a PTSD-specific database.

To address this, we are developing the comprehensive PTSD Biomarkers Database (PTSDDB), focusing on fluid-

based biomarkers as a resource for bringing together published findings within the context of study design and related results. Organizing and contextualizing published information and data are critical to understand the level of confirmatory and contradictory evidence for single biomarkers. Overall, this work represents one first step to crossing the translational divide between basic science discovery and clinical implementation. Considering findings for single biomarkers in tandem with details around study design may offer insights into the robustness and replicability of studies. Here, we present the first version of the PTSDDB biomarker database that provides a comprehensive and interactive view of results from an extensive, systematic curation effort in over 100 PTSD-focused articles. We aim to continually build on this resource in the future to enable a better interpretation of the state of the field in biomarker research that supports formal meta-analyses around single biomarkers in PTSD.

## Materials and Methods

### Curation procedure and database content

*Corpus selection* Articles included in the PTSDDB were compiled via two routes: recommendations and referrals from experts in the field and mining cited references from PTSD review publications. In general, publications included in this deep-dive database were original articles published from 1997 to 2017 that evaluated fluid-based biomarkers in humans, with a focus on PTSD patients vs control populations (e.g. healthy controls, trauma-exposed controls and/or patients with psychiatric disorders or other comorbidities). Exclusion criterion included publications that did not include a PTSD population, those that included PTSD patients but in the absence of fluid-based biomarkers, or that were preclinical studies.

*Data extraction and quality assessment* The biomarker metadata information contained in the resulting 109 publications was manually extracted independently by five independent trained curators and added to a data model template. The spectrum of curated metadata covers many fields, including study design, demographics, study findings, assay information and statistical methods. Ultimately, three rounds of quality control (QC) were conducted to ensure the fidelity of the metadata. In each round of QC, the metadata was reviewed by a distinct curator; if inconsistencies were found, curators worked together to reach a consensus. While there exist other QC procedures ensuring the quality of the curated data such as inter-curator agreement, these involve significant time constraints as they require two distinct curators working in parallel to extract the same

**Table 1.** Types of information extracted from each manuscript and stored as entries in the PTSDDDB. For each data category, extensive information was curated and stored as separate entries in the PTSDDDB. For example, the Data Category “Biomarker” includes information on Biomarker name, HUGO ID or another acronym, gene symbol/identifier, and biomarker application (e.g., biomarkers for disease risk, patient stratification, diagnostic marker, predictive markers of disease severity or treatment response, and safety/toxicity biomarkers).

| Model/data category                 | Fields   |
|-------------------------------------|--|
| Publication                         | PubMed identifier, authors information (e.g. names, year of publication and geographical information)  |
| Biomarker                           | Biomarker name, HUGO ID or other acronym, gene symbol, biomarker application. Protein, gene or miRNA biomarkers are coded using the HGNC nomenclature if possible. Small molecules nomenclatures are prioritized in the following order: ChEBI, PubChem and InChIKeys. |
| Time point                          | Study time point (e.g. 6 weeks post-trauma, 2 years post-trauma)   |
| Approach panel details              | Name, statistical method, cutoff used (e.g. <i>P</i> -value, False Discovery Rate, etc.), whether the biomarker was part of a panel, other analytes included, risk SNP, allele risk and additional notes on risk SNP   |
| Numerical summary                   | Statistics (i.e. mean and standard deviation [SD]) about the biomarker measurements for primary indication, trauma-exposed controls, other central nervous system and healthy controls   |
| Clinical instrument                 | Clinical instrument for primary indication, trauma in adult, childhood and lifetime (e.g. Clinician-Administered PTSD SCALE [CAPS], PTSD Checklist [PCL], Structured Clinical Interview for DSM-5 [SCID], etc.)  |
| Clinical study                      | Type of study (e.g. cross-sectional, longitudinal), timeline, challenge type, treatment response study, number of subjects per indication (e.g. trauma-exposed PTSD, trauma-exposed controls, healthy controls, other indications)                                     |
| Indication                          | Name and specifics of the condition (e.g. PTSD, childhood trauma, maternal PTSD)   |
| Comorbidity                         | Name, comorbidity measurements (e.g. mean and SD)  |
| Numerical readouts                  | Statistical details of each different group included in the study (e.g. mean, SD of the biomarker for primary indication or control)   |
| Inclusion/exclusion criteria        | The description of inclusion and exclusion criteria provided by each publication   |
| Cohort name and demographic details | Details about the cohort (e.g. percent female plus the overall mean and SD in trauma-exposed PTSD vs trauma-exposed and healthy controls, mean age and SD age of subjects)   |
| Ancestry                            | Ancestry details for Caucasian, African American and other ancestry of the cohort (e.g. percentage of each group included in the study)  |
| Study findings                      | Direction of change of the biomarker in cases vs controls as specified by the study authors (e.g. increased, no change, decreased), specific circuit changes, notes and descriptions   |
| Assay                               | Assay details: assay brand, probe, fluid, biological substrate, assay brand, assay limit detection and measurement units   |
| Assay calculations                  | Mean and SD concentration (in primary indication, trauma-exposed controls, healthy controls and CNS controls), sigma combined and effect size  |
| Statistical info                    | Mean, SD and variance, methylation change, <i>t</i> and <i>P</i> -value  |

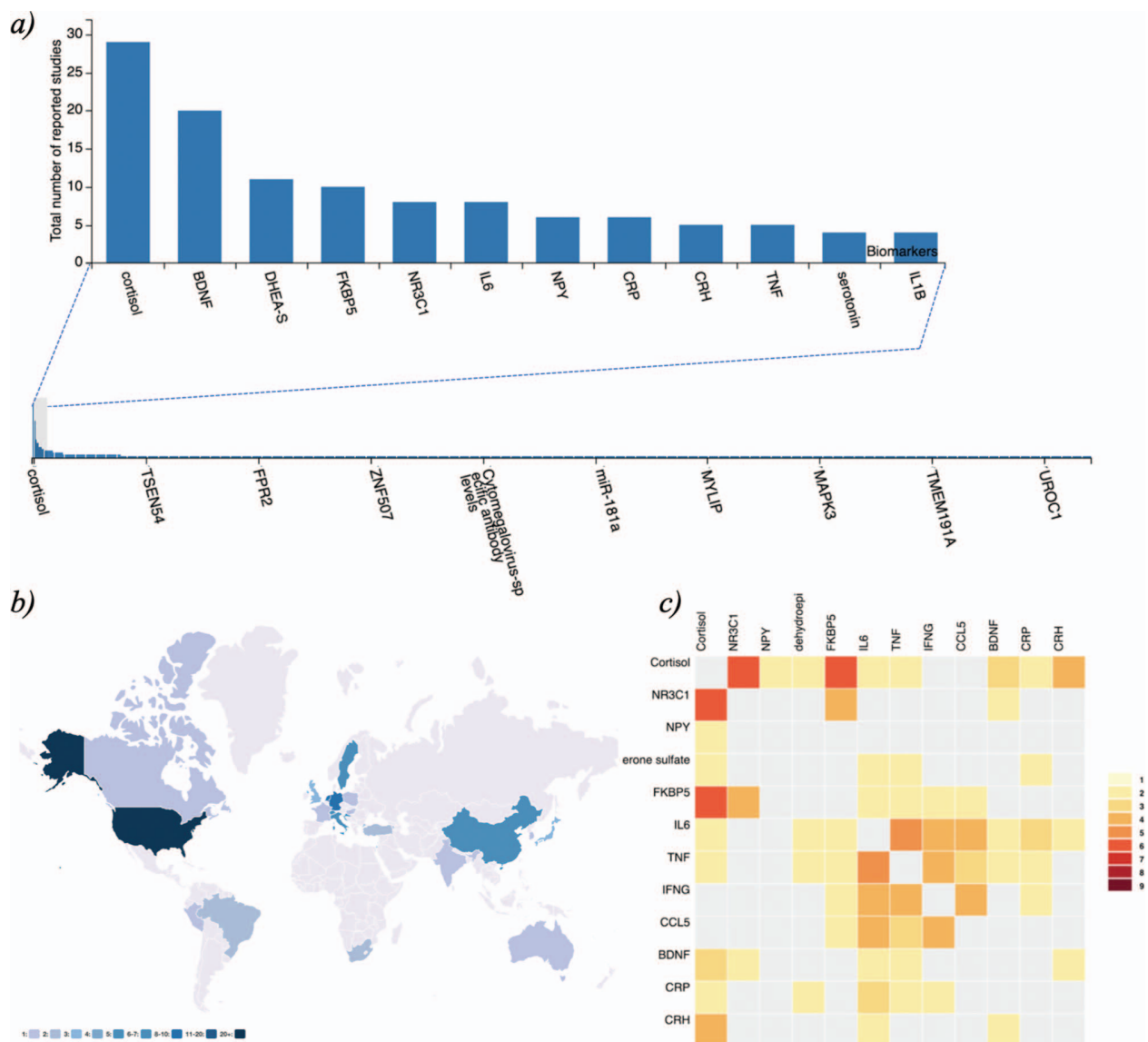
information. On the other hand, our QC procedure allowed us to include a larger amount of biomarker metadata while maintaining the quality of data curation. The data model template for metadata extraction was predefined based on the initial set of 10 articles. However, new metadata fields were added as necessary to accommodate new types of information being extracted from additional manuscripts.

### Database design and web application implementation

**Database model** One challenge in data integration when curating disparate sources of information is organizing this knowledge according to a common schema, which greatly influences subsequent steps of data management and analysis. In order to structure the information comprised

in the 193 columns of the curation template/worksheet, we designed a data model storing this information into 17 different models (e.g. publication, biomarker, clinical study, cohort metadata, etc.), which are represented as tables in the database (Table 1). Next, we implemented a parser of the curation template that populates the MySQL database and controls the quality of the worksheet by identifying duplicates, checking the syntax and normalizing terms. Finally, a web application integrated the database enabling users to query, visualize and analyze the curated content as illustrated in this manuscript.

**PTSDDDB implementation** The application was implemented following a model-view-controller (MVC) software architecture. The back-end is written in Python using the Django web framework technology (<https://www.djangoproject.com>).



**Figure 1.** PTSSDDM - Biomarker Data and Integrated Metadata: a) Frequency plot of biomarkers captured in the current version of the PTSSDDB, b) Geographical map displaying locations of institutions in the curated literature, and c) Heatmap visualization showing the frequency of individual biomarkers studied together in the same articles curated in the PTSSDDB. Descriptions of these visualizations are outlined in the Supplementary Information, and these figures can be dynamically explored at <https://ptsd.scai.fraunhofer.de/frequencies> and <https://ptsd.scai.fraunhofer.de/literature>.

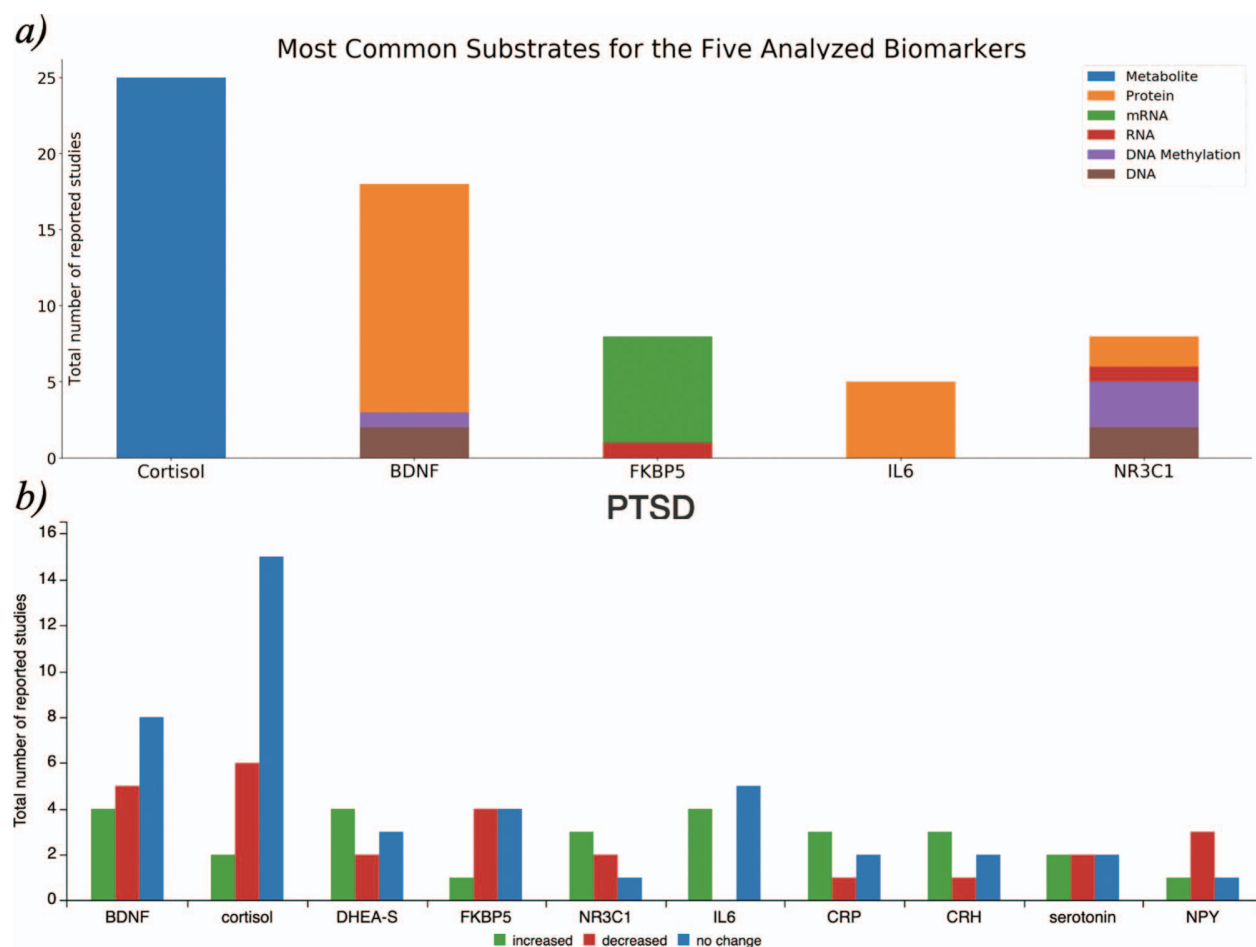
com/). Django embraces the MVC paradigm by storing the data into a MySQL relational database controlled by views that are responsible for querying the database and rendering its content to the users. The front-end renders interactive visualizations using a collection of powerful Javascript libraries: D3.js (<https://d3js.org/>), C3.js (<http://c3js.org/>), DataTables (<https://datatables.net/>) and DataMaps (<https://datamaps.github.io/>). Because the main goal of the web application is data exploration and visualization, the front-end is powered by Bootstrap, thereby retaining full compatibility with a broad range of devices (e.g. smartphones, tablets, laptops, etc.). Finally, PTSSDDB is

complemented with a RESTful API documented with an OpenAPI specification (<https://www.openapis.org>).

### Results and discussion

The PTSSDDB is an interactive database that catalogs information on more than 900 physiological markers, extracting data from over 100 manuscripts. The current PTSSDDB demonstrates our ability to successfully capture and organize large amounts of knowledge around PTSD physiological markers reported in the literature, using this information to support the creation of interactive data visualizations. By





**Figure 2.** a) Biological substrates of the five most frequently reported biomarkers in PTSDDB when they are studied as a metabolite, protein, or RNA. The source code to reproduce this figure is available at <https://github.com/ddomingof/PTSDDB-Resources>. b) Relative changes in the ten most common biomarkers captured in the database. This figure can be explored interactively at [https://ptsd.scai.fraunhofer.de/relative\\_changes](https://ptsd.scai.fraunhofer.de/relative_changes).

expanding on the PTSDDB in the future, we aim to enable the broad investigation of biomarkers implicated in PTSD pathogenesis.

### Biomarkers overview

The first page, ‘biomarkers overview’, presents an overview of the biomarker knowledge available in the database, depicted by the frequency of captured biomarkers across studies (Figure 1a), the biofluids in which they were measured, the relative changes reported and the biological substrates captured (e.g. DNA, RNA and protein; Figure 2a).

### Direction of change by biomarker

Also, the second page of the PTSDDB provides information on ‘direction of change by biomarker’ with visualizations that summarize the directionality of biomarker

findings (i.e. whether a biomarker was observed to be increased, decreased or unchanged in cases vs controls). For example, Figure 2b summarizes the reported changes in 10 of the most common biomarkers captured in this database. To better contextualize the direction of change, the functionality of this page allows users to assess changes based on the biological substrate, where the biomarker was measured, as well as a dedicated page to explore metadata information, which will subsequently be described.

The current version of the PTSDDB provides a ‘proof-of-concept’ that such visualizations can be successfully created and enable users to interact with articles and data curated in the PTSDDB. The next iteration of the PTSDDB will include a comprehensive landscape analysis of the PTSD biomarker literature so that these visualizations will facilitate researchers’ ability to investigate and critically evaluate the metadata information. For instance, users may glean important information by evaluating and comparing studies

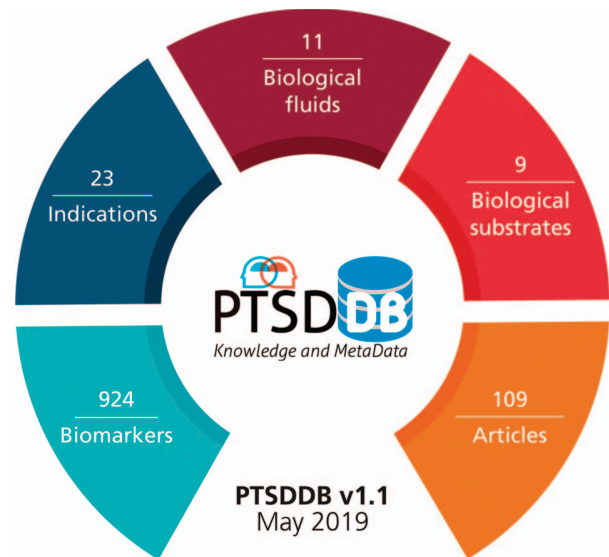
around factors that may impact reported findings, such as study type (cross-sectional vs longitudinal), sample size (e.g. number of subjects, controls, etc.), study population (military vs civilian, men vs women) and experimental methods (e.g. how was the biomarker measured).

### Study metadata

While abstracts and results sections summarize the essential findings in biomarker publications, capturing the level of evidence supporting single biomarkers requires information such as sample size or statistical measurements (e.g. mean, standard deviation, etc.), and these data are often unstructured in the form of figures or supplementary files. By storing and cataloging biomarker-related metadata, and then exposing it to the user via the ‘study metadata’ page, the PTSDDDB enables users to easily access and query this type of information. In contrast to the previously described information, this page focuses on the sparse world of clinical metadata, which is crucial to compare various study designs that are often complex and heterogeneous in nature. Here, users can first search for studies containing a particular biomarker and then inspect associated metadata (e.g. type of study, duration, challenge, trauma, sample size, assay, diagnostic criteria, etc.) for further analysis. Additionally, users can filter the studies by the specific application of the putative biomarker (i.e. diagnostic, prognostic, risk and stratification), allowing for more precise inquiries. Finally, a quick search box lists the biomarkers analyzed in a given study in order to facilitate the linkage between the meta information displayed in this page with the rest of the pages in PTSDDDB (e.g. ‘biomarkers overview’ or ‘direction of change by biomarker’), which are focused on providing a comprehensive overview of the results reported in the studies.

### Literature analysis

Biomarker research is often driven by current trends, technologies and specific hypotheses related to domain expertise. Currently, the increasing quantity of data, information and knowledge makes it incredibly complicated for researchers to stay abreast of all new studies published in a given area of interest. Thus, it is essential to provide researchers with an overview of what biomarkers have already been investigated as well as how the study was conducted. This information not only allows scientists to be aware of what has been studied but also may encourage collaboration among those working on similar hypotheses. To provide an overview of the literature included in the database and foster new research, PTSDDDB includes a page with novel visualizations, ‘literature analysis’, illus-



**Figure 3.** Content of the current version of PTSDDDB (May 2019): the database contains 109 articles, 924 biomarkers, 23 indications or distinct manifestations of PTSD, 11 biological fluids, and 9 substrates in which the biomarkers were tested.

trating which biomarkers are frequently studied together; where, how, and when were the studies were conducted; or in which biological substrates the biomarkers were measured. First, a table displays the main article information: PubMed-ID, title, journal, authors and year of publication. Second, map-based visualizations represent the geographical distribution of the analyzed articles to help identify PTSD-focused research hubs in the USA and across the globe, which may help identify collaborative opportunities for biomarker replication and validation (Figure 1b). Third, a histogram of the years when the articles included were published (Supplementary Information). Finally, two different heatmaps depict which biomarkers are frequently reported together in publications (Figure 1c) and which are studied in the same bio-fluid (Supplementary Information). By exploring this, we can investigate which combinations of biomarkers are most frequently studied in concert (e.g. cortisol and dehydroepiandrosterone sulfate cortisol and *NR3C1* were measured together in five of the studies that we have included so far in the database) or which biological substrates have been used together in clinical studies.

### Database content

Since PTSDDDB contains a large number of variables and metadata (Figure 3), it is an arduous task to implement interactive visualizations for every possible database query. Therefore, the last page, ‘database content’, contains a RESTful API (<https://ptsd.scai.fraunhofer.de/swagger-ui>) that exposes the database as well as a summary table of

the database, providing both interactive and programmatic interfaces to query, browse and navigate its content. The API is the gateway for researchers who are interested in data models that cannot be accessed through the interactive visualizations presented before (i.e. ancestry information, assay details and calculations, statistical information and inclusion/exclusion criteria). This enables researchers to access specific information extracted from the study, ranging from inclusion/exclusion criteria (e.g. type of medication excluded, comorbidities excluded, etc.) to details about the equipment used in the study (e.g. machine, brand, limits of detection, etc.). Furthermore, users can access associated statistics (e.g. means, standard deviations, *P*-values and fold changes calculated when comparing the biomarker in cases vs controls) in order to conduct or complement future meta-analyses. Finally, the API handles advanced database queries for extracting biomarker information that can be used to conduct complementary bioinformatics analyses as outlined by Zhang *et al.* in the context of colorectal cancer.

## Conclusion

The PTSDDDB organizes knowledge in the field of PTSD to provide a review of the literature, bringing together results from different studies so that researchers can evaluate results of single biomarkers, understand how they were measured, in what population and in what clinical contexts of use. This first version of the PTSDDDB involved significant curation and harmonization of information from disparate biomarker studies and related literature and storing this information in a database. In the future, we plan to integrate PTSDDDB into Brain Commons (<https://www.braincommons.org>), a big data cloud-based platform for computational discovery designed with user-friendly tools so that the PTSDDDB can be regularly updated and openly shared with the research community. In summary, to the best of our knowledge, the PTSDDDB is the first resource designed to catalog biomarker knowledge and metadata in PTSD and is complemented with a comprehensive web application that provides interactive visualizations and tools to query the cataloged knowledge. As this resource expands to capture all known knowledge around PTSD biomarkers, we aim to facilitate more formal meta-analyses so that robust conclusions may be drawn around the state of the field, in turn leading to new hypotheses for future studies.

## Supplementary data

Supplementary data are available at *Database* Online.

## Acknowledgements

We would like to thank the curators involved in the extraction of metadata from the articles included in the database.

## Funding

Cohen Veterans Bioscience, a 501(c)3 non-profit research organization.

*Conflict of interest.* A.P., H.L., K.D., N.D., L.L. and M.H. are employees of the non-profit funder of the research.

## References

1. American Psychiatric Association (2013) *Diagnostic and Statistical Manual of Mental Disorders*, 5th edn. DMS Library. <https://doi.org/10.1176/appi.books.9780890425596>.
2. Bina, R.W. and Langevin, J.P. (2018) Closed loop deep brain stimulation for PTSD, addiction, and disorders of affective facial interpretation: review and discussion of potential biomarkers and stimulation paradigms. *Front. Neurosci.*, **4**, 300. <https://doi.org/10.3389/fnins.2018.00300>.
3. Dai, H.J., Wu, J.C.Y., Lin, W.S. *et al.* (2014) LiverCancerMarker-RIF: a liver cancer biomarker interactive curation system combining text mining and expert annotations. *Database (Oxford)*, **ba085**. <https://doi.org/10.1093/database/bau085>.
4. Daskalakis, N.P., Cohen, H., Nievergelt, C.M. *et al.* (2016) New translational perspectives for blood-based biomarkers of PTSD: from glucocorticoid to immune mediators of stress susceptibility. *Exp. Neurol.*, **284**, 133–140. <https://doi.org/10.1016/j.expneurol.2016.07.024>.
5. Daskalakis, N.P., Provost, A.C., Hunter, R.G., *et al.* (2018) Non-coding RNAs: stress, glucocorticoids, and posttraumatic stress disorder. *Biol. Psychiatry*, **83**, 849–865. <https://doi.org/10.1016/j.biopsych.2018.01.009>.
6. Etkin, A. and Wager, T.D. (2007) Functional neuroimaging of anxiety: a meta-analysis of emotional processing in PTSD, social anxiety disorder, and specific phobia. *Am. J. Psychiatry*, **164**, 1476. <https://doi.org/10.1176/appi.ajp.2007.07030504>.
7. Hamosh, A., Scott, A.F., Amberger, J.S. *et al.* (2005) Online Mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **33**, D514–D517. <https://doi.org/10.1093/nar/gki033>.
8. Jovanovic, T. and Ressler, K.J. (2010) How the neurocircuitry and genetics of fear inhibition may inform our understanding of PTSD. *Am. J. Psychiatry*, **167**, 648–662. <https://doi.org/10.1176/appi.ajp.2009.09071074>.
9. Kang, J. *et al.* (2015) Peripheral biomarker candidates of post-traumatic stress disorder. *Exp. Neurobiol.*, **24**, 186–196. <https://doi.org/10.5607/en.2015.24.3.186>.
10. Kinoshita, J. and Clark, T. (2007) Alzforum. In: *Neuroinformatics*, pp. 365–381. [https://doi.org/10.1007/978-1-59745-520-6\\_19](https://doi.org/10.1007/978-1-59745-520-6_19)
11. Passos, J.C., Vasconcelos-Moreno, M.P., Costa, L.G. *et al.* (2015) Inflammatory markers in post-traumatic stress disorder: a systematic review, meta-analysis, and meta-regression. *Lancet Psychiatry*, **2**, 1002. [https://doi.org/10.1016/S2215-0366\(15\)00309-0](https://doi.org/10.1016/S2215-0366(15)00309-0).

12. Michopoulos,V., Norrholm,S.D. and Jovanovic,T. (2015) Diagnostic biomarkers for posttraumatic stress disorder: promising horizons from translational neuroscience research. *Biol. Psychiatry*, **1**, 344–353. <https://doi.org/10.1016/j.biopsych.2015.01.005>.
13. Pietrzak,R.H., Goldstein,R.B., Southwick,S.M. *et al.* (2012) Psychiatric comorbidity of full and partial posttraumatic stress disorder among older adults in the United States: results from wave 2 of the National Epidemiologic Survey on Alcohol and Related Conditions. *Am. J. Geriatr. Psychiatry*, **20**, 380–390. <https://doi.org/10.1097/JGP.0b013e31820d92e7>.
14. Pitman,R.K., Rasmusson,A.M., Koenen,K.C. *et al.* (2012) Biological studies of post-traumatic stress disorder. *Nat. Rev. Neurosci.*, **13**, 769–787. <https://doi.org/10.1038/nrn3339>.
15. Speer,K., Upton,D., Semple,S. *et al.* (2018) Systemic low-grade inflammation in post-traumatic stress disorder: a systematic review. *J. Inflamm. Res.*, **22**, 111–121. <https://doi.org/10.2147/JIR.S155903>.
16. Tamborero,D., Rubio-Perez,C., Deu-Pons,J. *et al.* (2018) Cancer genome interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Med.*, **10**, 25. <https://doi.org/10.1186/s13073-018-0531-8>.
17. Yang,I.S., Ryu,C., Cho,K.J. *et al.* (2007) IDBD: infectious disease biomarker database. *Nucleic Acids Res.*, **36**, D455–D460. <https://doi.org/10.1093/nar/gkm925>.
18. Yerlikaya,S., Broger,T., MacLean,E. *et al.* (2017) A tuberculosis biomarker database: the key to novel TB diagnostics. *Int. J. Infect. Dis.*, **56**, 253–257. <https://doi.org/10.1016/j.ijid.2017.01.025>.
19. Zhang,X., Sun,X.F., Cao,Y. *et al.* (2018) CBD: a biomarker database for colorectal cancer. *Database (Oxford)*, **2018**, bay046. <https://doi.org/10.1093/database/bay046>.
20. Zoladz,P.R. and Diamond,D.M. (2013) Current status on behavioral and biological markers of PTSD: a search for clarity in a conflicting literature. *Neurosci. Biobehav. Rev.*, **37**, 860–895. <https://doi.org/10.1016/j.neubiorev.2013.03.024>.

## Conclusions

This work crosses the translational divide between basic science discovery and clinical implementation by providing the first database in PTSD biomarker research. PTSDDDB provides a comprehensive overview of results from an extensive curation effort in over one hundred articles reporting findings on fluid-based biomarkers. Additionally, this resource is not presented as a static database, but complemented with visualizations and a sophisticated API designed to assist researchers in analyzing and elucidating the results of the biomarker studies incorporated. In summary, the presented work paves the way not only for the harmonization and unification of biomarker knowledge in the PTSD area, but also for conducting meta-analyses with each of the biomarkers in the database. This, in turn, could provide insights that explain the conflicting and inconsistent evidences reported in the scientific literature until now.



# 7

## Using multi-scale genetic, neuroimaging and clinical data for predicting Alzheimer's disease and reconstruction of relevant biological mechanisms

### Introduction

The question "can we detect Alzheimer's disease early enough to treat it in its early stages?" has prompted the development of numerous predictive models using large clinical datasets [127–129]. Although the prognostic power of such models could be higher than any clinician, they are presented as *black boxes* due to the lack of biological insights into how the underlying predictions are derived. However, bridging the gap between patient-level data and the mechanistic knowledge necessary to interpret the predictions of a model is still a challenging task. We propose a novel methodology to close this gap. Namely, we linked the features derived from a predictive model trained on the largest AD clinical study with NeuroMMSig, the knowledge-driven mechanistic inventory presented in chapter 5. This crosstalk between both approaches uncovered potential mechanisms driving the transition from normal and Mild Cognitive Impairment (MCI) cases to AD patients.

Reprinted with permission from "Shashank Khanna & Daniel Domingo-Fernández *et al.*. Using Multi-Scale Genetic, Neuroimaging and Clinical Data for Predicting Alzheimer's Disease and Reconstruction of Relevant Biological Mechanisms. *Scientific Reports*, Volume 8, Article number: 11173 **2018**". Copyright © Shashank Khanna & Daniel Domingo-Fernández 2018.



# SCIENTIFIC REPORTS

OPEN

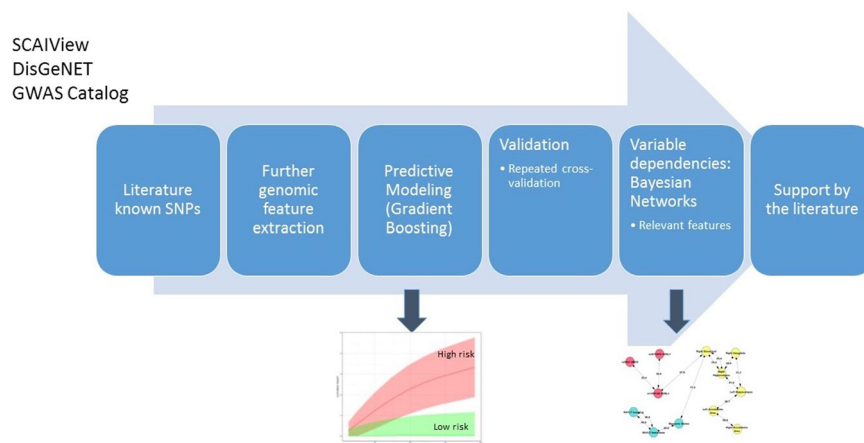
## Using Multi-Scale Genetic, Neuroimaging and Clinical Data for Predicting Alzheimer's Disease and Reconstruction of Relevant Biological Mechanisms

Shashank Khanna<sup>1,2</sup>, Daniel Domingo-Fernández<sup>1,2</sup>, Anandhi Iyappan<sup>1,2</sup>,  
 Mohammad Asif Emon<sup>1,2</sup>, Martin Hofmann-Apitius<sup>1,2</sup> & Holger Fröhlich<sup>1,2,3</sup>

Alzheimer's Disease (AD) is among the most frequent neuro-degenerative diseases. Early diagnosis is essential for successful disease management and chance to attenuate symptoms by disease modifying drugs. In the past, a number of cerebrospinal fluid (CSF), plasma and neuro-imaging based biomarkers have been proposed. Still, in current clinical practice, AD diagnosis cannot be made until the patient shows clear signs of cognitive decline, which can partially be attributed to the multi-factorial nature of AD. In this work, we integrated genotype information, neuro-imaging as well as clinical data (including neuro-psychological measures) from ~900 normal and mild cognitively impaired (MCI) individuals and developed a highly accurate machine learning model to predict the time until AD is diagnosed. We performed an in-depth investigation of the relevant baseline characteristics that contributed to the AD risk prediction. More specifically, we used Bayesian Networks to uncover the interplay across biological scales between neuro-psychological assessment scores, single genetic variants, pathways and neuro-imaging related features. Together with information extracted from the literature, this allowed us to partially reconstruct biological mechanisms that could play a role in the conversion of normal/MCI into AD pathology. This in turn may open the door to novel therapeutic options in the future.

Alzheimer's Disease (AD) is among the most frequent neuro-degenerative diseases in people above 65 and affects more than 45 Million people worldwide<sup>1</sup>. It is a chronic disease that usually starts slowly with a pre-symptomatic phase and worsens over time<sup>2</sup>. Early diagnosis is essential for successful disease management and chance to attenuate symptoms by disease modifying drugs. In the past, a number of cerebrospinal fluid (CSF), plasma and neuro-imaging based biomarkers have been proposed for that purpose<sup>3</sup>. Still, in current clinical practice, AD diagnosis cannot be made until the patient shows clear signs of cognitive decline, which can partially be attributed to the multi-factorial nature of AD<sup>4</sup>. AD pathology covers multiple biological scales, ranging from disease risk increasing genomic variants over altered intra-cellular signaling events and regional brain atrophy up to neuro-psychological behavior<sup>5</sup>. Hence, there is a need for establishing robust biomarker signatures covering multiple biological scales, which allow for early AD diagnosis. Several authors proposed models, which discriminate between AD and mild cognitively impaired (MCI) patients using subsets of markers from different data modalities<sup>5-9</sup>. A model to predict the time to conversion from 346 MCI into AD based on clinical data, neuro-imaging features and highly restricted genotype information (only 2 SNPs) was developed by Lee *et al.*<sup>10</sup>. The authors developed a Bayesian functional linear Cox model, which they evaluated based on simulation studies. Based on these simulations they reported an integrated area under ROC curve of 84%. In addition to the general limitation of such a purely simulation based validation the actual utility for clinical practice would have to be validated in

<sup>1</sup>Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Schloss Birlinghoven, Sankt Augustin, 53754, Germany. <sup>2</sup>Bonn-Aachen International Center for Information Technology (B-IT), University of Bonn, 53113, Bonn, Germany. <sup>3</sup>UCB Biosciences GmbH, Alfred-Nobel Str. 10, 40789, Monheim, Germany. Shashank Khanna and Daniel Domingo-Fernandez contributed equally to this work. Correspondence and requests for materials should be addressed to H.F. (email: [holger.froehlich@ucb.com](mailto:holger.froehlich@ucb.com))



**Figure 1.** Overall approach to analyze ADNI data.

a follow-up study first. Moreover, the biological mechanisms driving the MCI to AD conversion remain entirely unclear.

More recently, Li *et al.*<sup>11</sup> individually investigated different baseline cognitive, neuro-psychological and neuro-imaging scores to predict MCI to AD conversion. The authors employed univariate Cox models with covariate adjustments for age, gender, APOE4 status and education level. Using 6-fold cross-validation they reported a time dependent area under ROC curves of 68–81% for the respective scores. Once again, the biological mechanisms driving the MCI to AD conversion remain unclear.

The goal of this work was two-fold: First, our aim was to establish a multivariate, multi-modal predictive model for the time to AD conversion of normal/MCI patients and to identify most relevant prognostic features. Our model integrated rich genotype information (including newly developed SNP functional pathway impact scores), neuro-imaging (volume measurements of brain regions, PET scan results) as well as clinical data from 900 normal and MCI individuals extracted from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) (<http://adni.loni.usc.edu/>), a large scale observational study started in 2004 to evaluate the use of diverse types of biomarkers in clinical practice. A second aim of this work was to better understand the biological mechanisms driving the conversion of normal/MCI into AD pathology, which may ultimately open the door to novel therapeutic options. To this end, we employed a combination of data driven probabilistic and knowledge driven mechanistic approaches. More specifically, we used Bayesian Networks to uncover the interplay across biological scales between genetic variants, pathways, PET scan results and neuro-imaging related features. Together with manually curated cause-effect chains extracted from the literature, this allowed us to partially reconstruct biological mechanisms that could play a role in the conversion of normal/MCI into AD pathology.

## Results

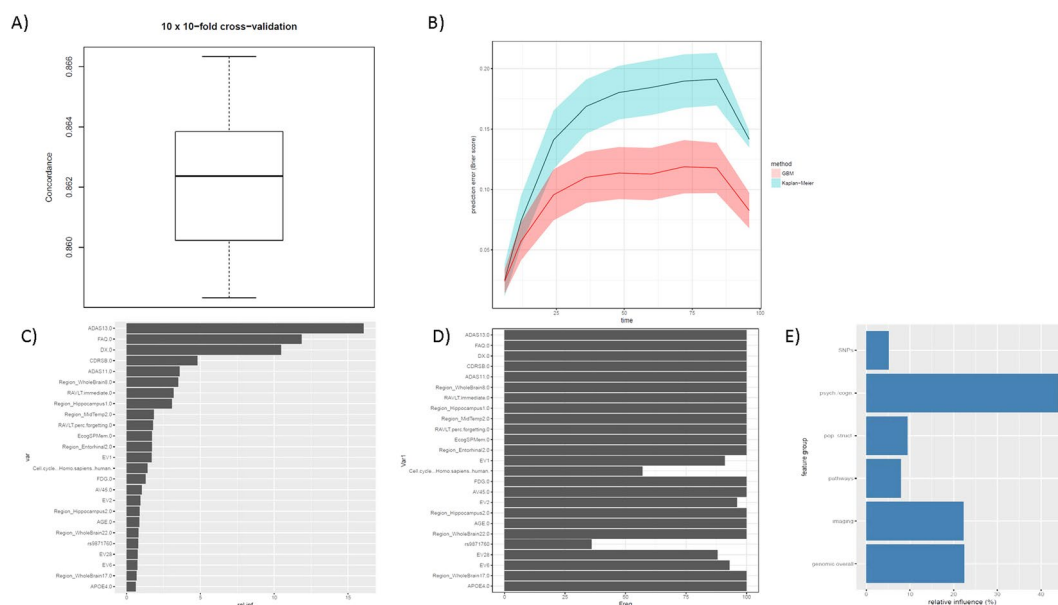
**Overview about Approach.** We extracted multi-modal baseline data from 315 normal and 609 mild cognitively impaired (MCI) patients from the ADNI database. 14 (4.4%) of the normal and 238 (39%) of the MCI patients developed AD during the 96 months of the study. We cannot exclude that patients without transition to AD pathology during study time developed AD later. Hence, their disease outcome has to be viewed as right censored.

The clinical baseline data of the altogether 924 patients used in this work comprised 73 variables with diagnosis, demographic information, age, gender, education level, neuro-psychological test, MRI and PET scan results, volume measurements of different brain regions as well as 300,000 single nucleotide polymorphisms (SNPs), which are commonly available from both ADNI1 and ADNI2/GO studies. Our overall approach to analyze these data and reduce their complexity contained six steps that are outlined in Fig. 1:

1. Literature mining of known disease associated SNPs (see Methods for details).
2. Further genomic feature extraction based on global population structure (principal components) plus a newly developed score to measure putative pathway impact of SNPs on individual patient level.
3. Development and evaluation of a predictive time-to-event model for normal/MCI to AD transition.
4. Estimation of (partially causal) dependencies between relevant features in the predictive model via Bayesian Network (BN) structure learning.
5. Validation of the BN with literature derived cause-effect relationships.

Steps 1. and 2. resulted into 313 pathway impact scores, 363 SNPs and 32 principal components that were added to the above mentioned 73 clinical baseline variables.

**Predicting the Time Dependent Alzheimer’s Disease Risk and Identification of Associated Relevant Baseline Characteristics.** *AD Can Be Predicted Accurately.* We developed an approach to appropriately integrate the multi-modal data used in this work within a machine learning framework to predict the AD risk for MCI and pre-symptomatic patients (see Methods for details). Our approach uses a weighted ensemble of



**Figure 2.** (A) Boxplot of cross-validated concordance index. (B) Prediction error (Brier score) as a function of time for GBM vs. Kaplan-Meier estimator. The prediction error curve is calculated on held out test data during the 10 times repeated 10-fold cross-validation procedure. The solid curve corresponds to the mean and the shaded area to the standard deviation. (C) 25 most relevant features according to GBM model trained on the whole tuning dataset. (D) Selection frequency of these features during the 10 times repeated 10-fold cross-validation procedure. (E) cumulative relative influence of feature groups in final model.

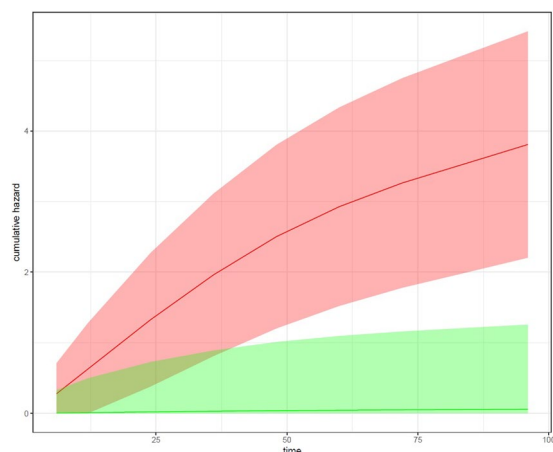
constraint decision trees - a Gradient Boosting Machine (GBM<sup>12</sup>) - to combine most relevant pathways, SNPs, principal components and clinical baseline variables into a final patient specific prediction score. GBM is an established machine learning method that is - due to its nature as an ensemble of decision trees - well suited to integrate heterogeneous data types (e.g. clinical features plus SNPs) on rather different numerical scales, as in our application<sup>13</sup>. Moreover, GBM allow for an embedded subset selection of most relevant features and appropriately deal with missing values in clinical data.

The prediction performance of our developed GBM based algorithm was assessed via a 10 times repeated 10-fold cross-validation procedure: Cross-validation randomly splits the overall data into  $k$  (here 10) folds, while successively one of these folds is left out for model validation and the rest for model training. The ability to predict the time to first AD diagnosis for patients in the validation set of each of the 10 GBM models was assessed via Harrell's concordance/C-index<sup>14</sup>, which is a generalization of the area under ROC curve measure used in binary classification. The C-index ranges from 0 to 1, where 0.5 indicates chance level.

As indicated in Fig. 2A our algorithm achieved a high prediction performance with a cross-validated C-index of 0.86. Figure 2B depicts the time dependent prediction error (in terms of Brier score) of the GBM model on held out test data during the repeated cross-validation procedure in comparison to a Kaplan-Meier estimator, showing clearly superior performance with low prediction error. We thus conclude that our employed multi-scale data allows for an accurate prediction of the time dependent risk to convert from normal/MCI to AD pathology. Notably, our developed GBM model achieved a significantly higher cross-validated C-index than another and popular ensemble based decision tree technique, Random Survival Forest<sup>15</sup>, elastic net penalized Cox regression<sup>16,17</sup> and two different Canonical Correlation Analysis (CCA) based methods<sup>18,19</sup> followed by conventional Cox regression, see Methods.

**Most Relevant Features are Interpretable.** To better understand the contribution of individual features for the AD risk prediction we ranked variables according to their relative importance in a final GBM model that was trained on all available data (Fig. 2C). The top 25 most relevant variables comprised, besides baseline diagnosis (DX), results of different neuro-psychological/cognitive assessments (Alzheimer's Disease Assessment Scale Cognitive Plus - ADAS13, ADAS11, Functional Assessment Questionnaire - FAQ, Clinical Dementia Rating - CDRSB, Rey Auditory Verbal Learning Test - RAVLT, Everyday Cognition Study Partner Report - EcogSPMem), neuro-imaging features (Region\_Hippocampus, Region\_Enthorhinal, Region\_MidTemp, Region\_WholeBrain8), PET and FDG PET imaging diagnosis (AV45, FDG), APOE4 status as well as patient age. Furthermore, different features describing the genetic population sub-structure (EV1, EV2,...) as well as the SNP functional impact on cell cycle were contained. It has been suggested that dysfunction in neuronal cell cycle reentry plays a fundamental role in AD pathology<sup>20</sup>. More specifically, the hypothesis has been stated that the disease is caused by aberrant re-entry of different neuronal populations into the cell division cycle<sup>21</sup>.

Notably, most of the top 25 were selected highly stable during the 10 times repeated 10-fold cross-validation procedure (Fig. 2D). That means the vast majority of GBM models trained during the cross-validation procedure contained the same most relevant features. This finding specifically includes the above mentioned cell cycle.



**Figure 3.** Cumulative hazard as a function of time for the 10% patients with highest AD risk scores (red) and 10% patients with lowest AD risk scores (green). Depicted are the average risk curves plus standard errors as confidence bands.

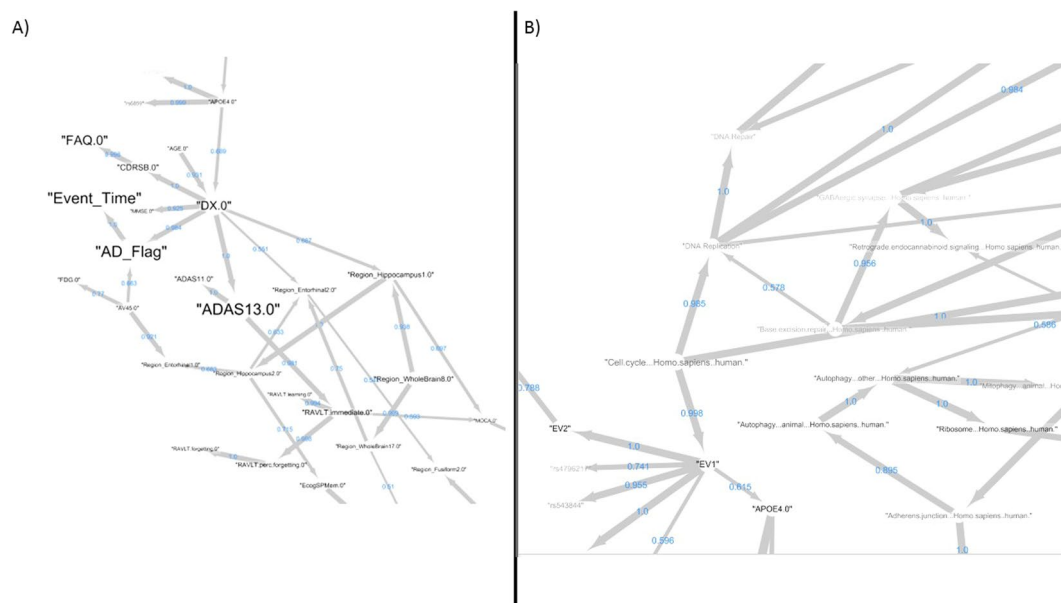
Altogether there were 170 features that were selected at least in 50 out of 100 times (see full list in Supplementary material). These features contained the neuro-psychological assessments (ADAS, Ecog, RAVLT, CDRSB, MMSE, FAQ), PET scanning results (AV45, FDG), APOE4 status, age, baseline diagnosis, educational status as well as different brain regions and pathways (including cell cycle). The most stably selected SNP rs10509663 (selected 70/100 times) has been associated with CSF levels of amyloid- $\beta$ . Misfolding of this peptide is a well known hallmark of AD that results into the characteristic plaques in the brain of AD patients<sup>22</sup>. Interestingly, immune system and ribosome were found as most stably selected pathways (84/100 times). It has recently been indicated that activation of the innate immune system plays a crucial role in disease progression<sup>23</sup>. Ribosome dysfunction has been observed as an early event in AD development<sup>24</sup>.

The most influential SNP in the final GBM model was rs9871760 (selected 36/100 times), which has been associated to the whole brain volume<sup>25</sup>. The TT or CT genotypes of the second most relevant SNP rs3756577 (CAMK2A, selected 32/100) have been associated with a nearly 8 times risk reduction for AD<sup>26</sup>. Two other examples include rs4263408 (selected 32/100 times) and rs6859 (selected 60/100 times). The SNP rs4263408 (UBE2K) has been found to affect amyloid- $\beta$  concentrations<sup>27</sup>. The SNP rs6859 (NECTIN2) has been associated with late AD onset<sup>28</sup>.

Altogether the cumulative relative influence of all genomically derived features (including APOE4 status) was ~22% in our model, and 109/170 features that were selected at least 50/100 times during the repeated cross-validation procedure were genomically derived. Figure 2E systematically visualizes the cumulative relative influence of different feature groups, such as SNPs, neuro-psychological/cognitive tests, features describing the genomic population sub-structure (principal components), SNP impact on pathways and neuro-imaging features. This demonstrates an equal contribution of neuro-imaging and genomic features, whereas neuro-psychological/cognitive test results have an almost twice as high cumulative influence.

**The Model Allows for Patient Stratification.** Figure 3 exemplifies the possibility to stratify patients by the predictions made by our model into “high risk” and “low risk” groups. More specifically the Figure compares the cumulative risk curves of 93 patients in the upper 10% quantile of the risk score produced by our model with 93 patients in the lower 10% quantile of the risk score. Both curves show a clear difference ( $p \approx 0$ , log rank test). We performed univariate statistical tests (Wilcoxon for continuous and  $\chi^2$ -test for discrete variables) for each individual feature used in our GBM model to better understand differences between the high risk and low risk group. P-values were corrected for multiple testing using the Benjamin-Yekutieli false discovery rate (FDR) control under dependency<sup>29</sup>. Accordingly, we found clear differences in the APOE4 status ( $FDR < 1e-9$ ), in all neuro-psychological assessment scores, PET imaging diagnosis ( $FDR < 1e-4$ ), rs405509 ( $FDR < 0.05$ ), ErbB signaling and olfactory transduction (both  $FDR < 0.05$ ). In the low risk group 63% of the patients were diagnosed as healthy at baseline, whereas in the high risk group all patients were already late phase mild cognitively impaired. According to dbSNP<sup>30</sup>, rs405509 is located in the APOE4 gene region and synergizes with the APOE4  $\epsilon 4$  allele in the impairment of cognition. The T allele has been identified as a risk factor for AD<sup>31</sup>. ErbB signaling and olfactory transduction both showed a significant difference in SNP pathway scores. However, the difference in the impact score was in both cases less than 1%. Hence, any further interpretation should be taken with care. However, we like to mention that olfactory dysfunction and insufficient ErbB signaling have both been associated with AD<sup>32,33</sup>.

**Bayesian Network Modeling Reveals Dependencies Between Relevant Features.** Our predictive GBM model altogether contained a set of 335 features. To gain a better understanding of the complex and multiple interactions between these features we developed a Bayesian Network (BN) model<sup>34</sup>. BNs belong to the family of probabilistic graphical models and enable the description of a complex multivariate distribution with many variables (here all relevant features in our predictive model). BNs can be visualized as graphs, where nodes



**Figure 4.** Edges appearing in more than 50% of 1000 Bayesian Network reconstructions based on random sub-samples of the data. Line thickness is proportional to the relative frequency of observing an edge in the 1000 network reconstructions, and the corresponding number is shown as edge label. The node size is proportional to the relative influence of the variable in the final GBM model, and the color reflects the selection frequency in the repeated cross-validation procedure (more black = higher stability). Sub-figures (A) and (B) depict two examples zooms into the overall network.

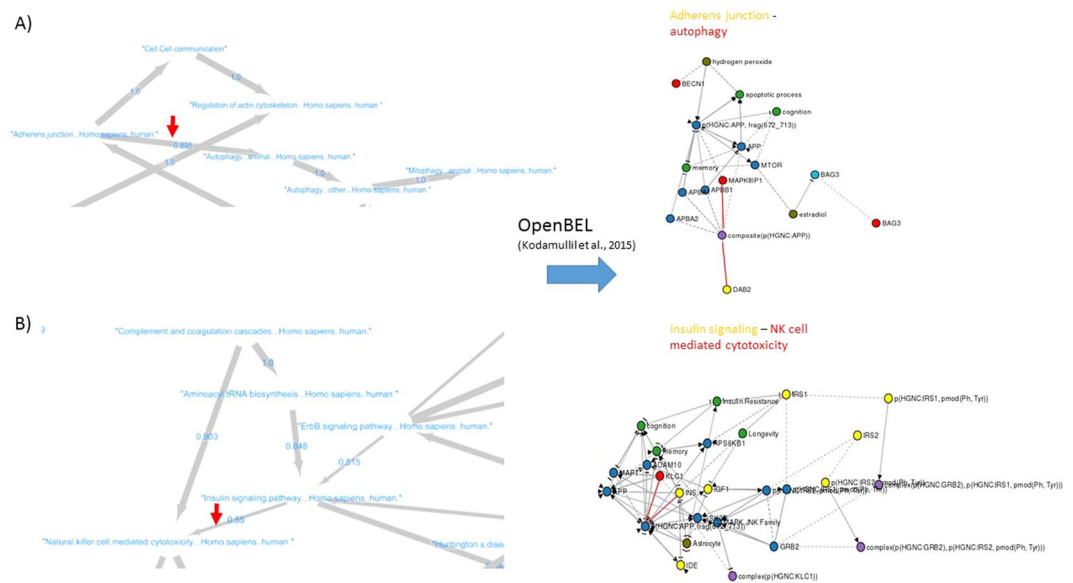
correspond to random variables and edges reflect conditional statistical dependencies. BNs have a long tradition in systems biology for learning and describing biological pathways<sup>35–37</sup>, because they - at least partially - allow for discovery of causal relationships from observed data (see Methods).

We developed a BN for the same 924 patients used in our final predictive GBM model. Importantly, we included the time until AD diagnosis together with a censoring indicator as variables. Six different BN learning algorithms were compared via a 10-fold cross-validation procedure (see Methods) and the best performing one (tabu search<sup>38</sup>) selected. Subsequently, we applied a non-parametric bootstrap to network learning, resulting into 257 edges appearing in more than 50% of 1000 network reconstructions based on random sub-samples of the data (see details in Methods section). Figure 4 shows two zooms into the network of these stable edges highlighting the direct dependency of the clinical outcome (Event\_Time, AD\_Flag) on baseline diagnosis (DX) and PET scan (AV45). PET scanning results manifest in the entorhinal region, which is known to be affected by AD pathology<sup>39</sup>. Baseline diagnosis is dependent on age and APOE4 mutation status. APOE polymorphic alleles are one of the major AD risk factors<sup>40</sup>. Baseline diagnosis influences neuro-psychological assessments (ADAS13, ADAS11, MMSE) and manifests in the entorhinal region and hippocampus, which is vulnerable specifically in early AD stages<sup>41</sup>.

Figure 4B further highlights the dependency of APOE4 status on sub-population structure (EV1), which is reflected by differences in cell cycle, hence supporting the above cited neuronal cell cycle hypothesis as one of the possible disease causes. Cell cycle includes DNA replication and repair, which is mirrored by a corresponding edge in our BN. DNA damaging by oxidative stress has been reported as one of the earliest detectable events in the progression to dementia<sup>42</sup>. More specifically, altered DNA repair has been observed in GABAergic neurons<sup>43</sup> and led to the idea of a therapeutic modulation of the GABAergic system in early AD stages<sup>44</sup>. GABAergic neurons distinctly express  $CB_1$  receptors, thus explaining the subsequent link to the endocannabinoid system<sup>45</sup>. Targeting this system has been discussed as a therapeutic option<sup>45</sup>.

To specifically validate some of the less obvious dependencies between pathways that were reflected via stable edges in our BN we checked the overlap of genes that could be mapped to the respective pathways based on KEGG<sup>46</sup> and Reactome<sup>47</sup> databases. The statistical significance of overlaps was assessed via a hyper-geometric test, and p-values corrected for multiple testing via the Benjamini-Yekutieli false discovery rate (FDR) under dependency<sup>29</sup>. Accordingly we obtained significant results ( $FDR < 5\%$ ) for 63/76 (83%) pathway pairs in our BN (see results in Supplements).

**Pathway Dependencies are Interpretable via Biological Mechanisms.** *Mapping of Stable Edges to Causal Biological Mechanisms.* To further validate pathway dependencies found by our BN methodology and to gain additional insights into underlying molecular mechanisms we employed a literature derived mechanistic AD disease model encoded in the OpenBEL language<sup>48</sup>. Briefly, this model describes cause-effect relationships between different biological entities, such as genes, SNPs and biological processes (e.g. neuronal death) in a purely qualitative manner. We developed an algorithmic approach to map nodes and edges in our network to the



**Figure 5.** Two examples of mapping stable BN edges to biological mechanisms via the OpenBEL AD graph by Kodamullil *et al.*<sup>48</sup>: (A) adherens junction and autophagy; (B) insulin signaling and natural killer cell mediated cytotoxicity. Biological entities mapping to the source of the edge marked by an arrow on the left hand side of the Figure are drawn in yellow in the OpenBEL graph on the right hand side. Biological entities mapping to the sink of the edge are shown in red. Red edges highlight the shortest among all possible paths connecting yellow and red nodes.

OpenBEL AD disease model. In conclusion, we could identify 12 cause-effect-relationship networks that could be linked to specific stable pathway-pathway edges in our BN (see Figures in Supplements). We have developed a software tool to explore our BN and associated mechanism mappings in a fully interactive manner. The tool is accessible under <http://neurommsig.scai.fraunhofer.de/bayesian>. In the following we discuss two selected mechanism mappings in greater detail.

**Example 1: Adherens Junction and Autophagy.** As a first example, our Bayesian Network predicts an association between adherens junction and autophagy, which has recently also been proposed in Nighot *et al.*<sup>49</sup> (Fig. 5A). Adherens junctions, also known as tight junctions, are comprised of epithelial cells that are present in all tissues, particularly in junction between cerebral epithelial cells of the blood-brain barrier (BBB)<sup>50,51</sup>. The BBB is a biochemical barrier which regulates the entry of blood based molecules into the brain and helps in maintaining ionic homeostasis within the brain. These barriers further inhibit diffusion of cellular components thereby protecting the central nervous system<sup>52,53</sup>. However, during pathological conditions such as AD, the BBB are disrupted increasing the cell permeability as well as accumulation of amyloid- $\beta$  resulting in autophagy.

Mapping of the edge between adherens junction and autophagy in the BN to OpenBEL encoded mechanisms allowed us to identify molecular players, which may play a role in the normal/MCI to AD transition: Proteolytic processing of amyloid precursor protein (APP) is one of the hallmarks of AD pathophysiology<sup>22</sup>. The processing of APP to amyloid- $\beta$  is greatly affected by the sub-cellular localization of  $\beta$  and  $\gamma$  secretases due to trans-membrane receptors as well as adapter proteins<sup>54,55</sup>. Internalization of APP via adapter proteins such as DAB2 triggers clathrin mediated endocytosis by binding to the YXNPXY motif region of APP triggering endocytosis<sup>56,57</sup>. Apart from DAB2, there are other adapter proteins that trigger the production of amyloid- $\beta$  namely APBA2 and APBA1. These two proteins are enriched in neurons and contain a phosphotyrosine binding site (PTB) domain<sup>58,59</sup>. These proteins are involved in cellular activities pertaining to neuronal transport and synaptic function. Unlike DAB2, APBA1 and APBA2 interact with the YENPTY motif region of APP and thereby affecting APP trafficking. During AD pathology, APBA1 protein modulates the secretory and endocytic trafficking of APP whereas APBA2 accelerates APP endocytosis which leads to autophagosomes that enhances amyloid- $\beta$  internalization<sup>60,61</sup>.

Autophagosomes are structures that facilitate the break down of accumulated amyloid- $\beta$  peptides by fusion with lysosomes. Lysosomes contain enzymes that break down accumulated peptides<sup>62,63</sup>. However, during AD progression, autophagosomes accumulate within neurons of AD patients. The scaffolding protein MAPK8IP1 is a regulator of autophagosomal motility by activating the c-Jun-N-terminal kinase (JNK), which mediates the JNK signaling cascade. JNK signaling formulates the formation of neurofibrillary tangles through direct phosphorylation of tau proteins further resulting in stress induced apoptosis in neuronal cells<sup>64,65</sup>. Furthermore, the activation of JNK signaling induces phosphorylation of Bcl-2 releasing beclin-1 protein which further aggravates autophagosome formation, resulting in cognitive decline<sup>66,67</sup>.

**Example 2: Insulin Signaling and Natural Killer (NK) Cell Mediated Cytotoxicity.** Another example is the BN predicted link between insulin signaling and natural killer (NK) cell mediated cytotoxicity (Fig. 5B), which has again been proposed in the literature<sup>68</sup>. Our extracted mechanism shows, how the axonal transport and APP trafficking may influence AD development: APP proteins are transported to distinct nerve cells through axons via anterograde pathways for maintaining homeostasis and neuronal function<sup>69</sup>. The fast anterograde transport is mediated through APP and Kinesin 1 (KLC1) and Fe65, adapter protein. The proteolytic processing of amyloid- $\beta$  occurs within the axons and through this process amyloid- $\beta$  is generated releasing the complex KLC1 from APP. During AD progression, the excessive production of amyloid- $\beta$  prevents the release of Kinesin and thereby restricts the axonal transport. The arrested axonal transport also triggers the phosphorylation of APP through the amyloidogenic pathway concomitantly releases the Fe65 and translocates into the nucleus to regulate the expression of stress-related genes including glycogen synthase kinase 3 beta (GSK3B)<sup>70,71</sup>.

Apart from APP and KLC1, insulin and IGF regulate neuronal stem cell activation, synaptic maintenance and neuroprotection<sup>72,73</sup>. Insulin regulates the glucose and lipid metabolism in the brain and thereby contributes to learning and memory<sup>74</sup>. It is known that insulin is locally produced in the brain and can be easily transported through the BBB<sup>75-77</sup>. The glucose metabolism is mediated by binding of the IGF to its receptor promoting the phosphorylation of the tyrosine residue and further phosphorylating the insulin receptor substrate (IRS) at the tyrosine residue. The two receptors, IRS-1 and IRS-2 are mediators of insulin-dependent mitogenesis and regulation of glucose metabolism, which is a part of the insulin-signaling pathway<sup>78,79</sup>. The phosphorylation of IRS1 results in the downstream activation of AKT, mammalian target of rapamycin (mTOR), growth receptor binding protein 2 (GRB2), mitogen-activated protein kinase (MAPK) and GSK3B, thereby promoting the APP transport and clearance of amyloid- $\beta$  from the BBB<sup>80,81</sup>. During AD progression, the insulin signaling pathway shows aberrant activity, resulting in increased accumulation of amyloid- $\beta$ , tau phosphorylation and decreased cerebral blood flow. Furthermore, the binding of IRS to its receptor is inhibited, resulting in decreased glucose metabolism and cognition<sup>82</sup>. Brain insulin resistance thus contributes to AD, a complex phenomenon accompanied by IGF-1 resistance and dysfunction of IRS-1 triggered by amyloid- $\beta$  oligomers, stimulating cognitive decline independent of AD pathology<sup>83</sup>.

## Discussion

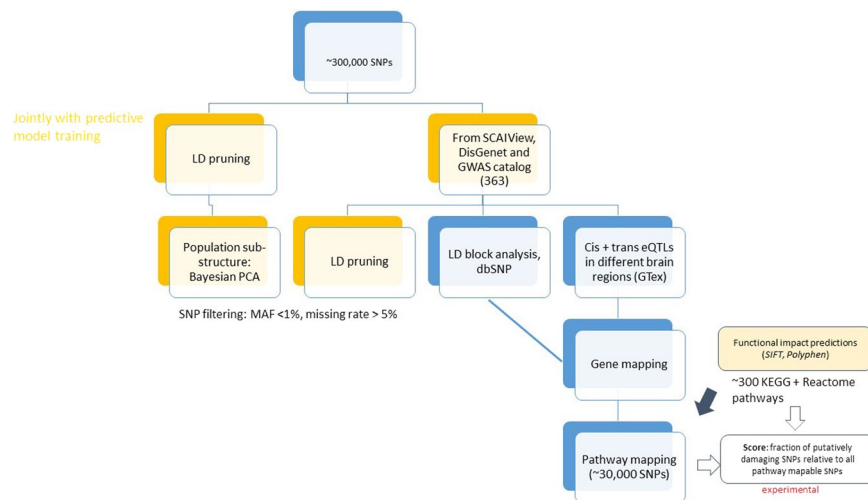
Determining the risk of an individual to develop AD is an important aspect to start treatment with disease modifying drugs as early as possible and to better manage the disease. To better address this need we developed a highly predictive time-to-event model for normal/MCI to AD transition based on multi-modal data from ADNI. For this purpose we proposed a novel approach to capture the functional impact of SNPs on pathway level for each individual patient. Analysis of our model confirmed the significant impact of the baseline diagnosis (cognitively normal, early or late stage mild impaired) for predictions and demonstrated the crucial role for stratifying patients into high and low risk groups. Further relevant features of our model include neuro-psychological assessment scores, neuro-imaging features (including PET scan results), age as well as genetic predisposition, which altogether contributed 22% cumulative influence. In addition to well known risk factors such as APOE4 status we specifically identified SNP functional impact on cell cycle as a relevant feature in our model, which agrees well with the hypothesis of AD being caused by dysfunction of the neuronal cell cycle reentry<sup>20</sup>.

As a further contribution of our work we tried to better understand dependencies of relevant features via a Bayesian Network model. Thanks to our proposed pathway functional impact score and with the help of a specifically developed algorithm we were able to relate several non-obvious links to detailed biological mechanisms, which constitutes a partial literature based validation. As two examples we discussed the mechanisms linking tight junction and autophagy as well as insulin signaling and NK cell mediated cytotoxicity in more detail and provided literature based evidence for the involvement into AD pathology. Altogether stable edges found in our BN provide a broad overview about the complex interplay of different AD risk factors and provide insights into their underlying biological mechanisms. Such insights could potentially help in developing novel and more mechanistic therapies in the future, which are critically needed in the field.

Of course, there are limitations of our work that we would like to mention: The whole work presented here was based on the ADNI cohort. Since this patient group primarily represents an amnesic rather than an epidemiologically selected population we cannot exclude population biases. Hence, a confirmation of our findings based on a different study cohort has to be conducted in the future. Furthermore, it is believed that AD pathology starts decades before actual diagnosis<sup>84</sup>. The age range of ADNI subjects is thus probably too late to find very early disease indications, and the follow-up time of 96 months is likely too short for many patients that were initially cognitively normal to allow for a definite AD diagnosis till end of study. Taking these aspects into consideration we therefore see the need for long lasting studies in a more epidemic population in the future. Key findings from our work could help designing such a study by identifying relevant factors to measure.

## Methods

**Clinical and Genomic Feature Extraction from ADNI.** *Data Preprocessing.* Clinical variables (including age, gender, education level, neuro-psychological assessments, pre-computed volume measurements of different brain regions and PET scan results) from all ADNI studies (ADNI1, ADNI2, ADNI-GO) were retrieved via the ADNImerge R-package (<https://adni.bitbucket.io/>), altogether comprising 73 features at study baseline after dropping variables with more than 65% missing values. For the 818 subjects in the ADNI1 study population 620,901 SNP calls were available via the Illumina Human610-Quad BeadChip platform. Further genomic data (730,525 SNPs, Illumina HumanOmniExpress BeadChip) was available for 432 subjects in ADNI2/GO. 979 patients were diagnosed as either normal or MCI at baseline, and 314,134 SNPs were found in common between both ADNI phases. Following common convention we encoded SNPs by the occurrence of a minor allele (0, 1, 2) while taking dbSNP as ref.<sup>85</sup>.



**Figure 6.** Approach to genomic feature extraction.

The initial set of 979 patients was reduced to 926 by filtering out individuals with kinship coefficient  $<0.1$  and inbreeding coefficient  $>0.1$ . Kinship coefficient was calculated by the PLINK method of moment for identity-by-descent analysis<sup>86</sup>. Inbreeding coefficient estimation was based on the method described in Yang *et al.*<sup>87</sup>. For both analyses we employed the SNPRelate software<sup>88</sup>. SNPs with MAF  $<1\%$  or missing rate  $>5\%$  were filtered out.

**Literature Derived SNPs.** We used the text mining software SCAIView<sup>89</sup>, the DisGeNET database<sup>90</sup> and GWAS catalogue<sup>91</sup> with a similar search query (“Alzheimer’s Disease” and “Homo sapiens”) to retrieve 1,866 putatively disease associated SNPs, of which 363 intersected with the 314,134 SNPs measured in both ADNI phases (Fig. 6). Application of LD pruning ( $r^2 < 0.2$ ) using SNPRelate reduced this number further to  $\sim 300$ . Notably, this step was done as part of predictive model training and more specifically also within the repeated cross-validation procedure.

**Principal Components.** In addition to knowledge derived, LD pruned SNPs, we considered the global genetic population structure by retrieving the top principal components (based on all 314,134 SNPs) via a Bayesian principal component analysis (PCA)<sup>92</sup> after LD pruning. We favored Bayesian PCA over conventional maximum likelihood based PCA here due to the high dimensionality of SNPs compared to the number of patients. Bayesian PCA effectively regularizes PCA and thus results into more stable and robust estimates. Again, Bayesian PCA was done as part of model training and thus within the repeated cross-validation procedure. Note that the extremely high dimensionality of SNP data prevents us from extracting all principal components due to prohibitive computational costs. However, extraction of the top  $k$  (here: 32, the default in SNPRelate) principal components can be done efficiently using Lanczo’s method<sup>93</sup>. We relied on the implementation provided in the SNPRelate software<sup>88</sup>. The proportion of explained variance by the top principal components was typically around 5–6%, depending on the actual set of patients in the training set. This fact highlights that the top principal components for sure do not describe the global population structure entirely. However, they may still capture useful signals for our predictive model. The fact that several eigenvectors were among the most relevant features (Fig. 2) in our final model supports this thought.

**SNP Based Pathway Impact Scores.** We performed a gene mapping of SNPs. This was done by

1. Considering all 363 putatively disease associated SNPs taken from the literature and all those in strong LD ( $r^2 > 0.8$ ). Each SNP in an LD block was then mapped to its closest genes via dbSNP<sup>30</sup>.
2. Considering significant (false discovery rate  $<5\%$ ) cis- and trans-eQTLs in different brain regions from GTex<sup>94</sup>. For this step the same SNPs as in the previous one were used.

Note that both steps can result into a mapping of one SNP to several genes. The entire set of genes was then extended to a pathway mapping, where pathways were taken from KEGG<sup>46</sup> and Reactome<sup>47</sup>. Around 30,000 SNPs were map-able to pathways altogether. For these  $\sim 30,000$  SNPs we used functional impact predictions from SIFT<sup>95</sup> and Polyphen<sup>2</sup><sup>96</sup>. We then developed an experimental score to capture the overall functional impact on each pathway per patient: This score was defined as the fraction of at least possibly damaging/deleterious SNPs relative to all pathway map-able SNPs. This score is a number between 0 and 1 for each individual patient and pathway.

**Multi-Modal Time-to-AD-Diagnosis Prediction. Intermediate Data Fusion with Gradient Boosting Machines.** The employed data is characterized by a high heterogeneity with respect to different statistical distributions and numerical scales (e.g. SNPs vs. neuro-imaging features). Gradient Boosting Machines (GBM) have



been introduced as a decision tree based ensemble learning technique that enables non-linear and non-parametric time-to-event prediction based on such heterogeneous data<sup>12</sup>. A GBM constitutes a weighted ensemble of weak decision tree classifiers (base learners) with restricted maximal depth (here 3). A higher maximal tree depth results into more complex base learners, that capture higher order interactions between variables (here: up to 3-way interactions). On the other hand, a tree depth of 1 corresponds to simple decision stumps and can require longer boosting, depending on the overall optimal complexity of the GBM model. The reason is that the actual number of trees in the ensemble (and thus overall complexity of the GBM model) critically depends on the number of boosting steps, which is a tunable hyper-parameter. Depending on the maximal depth and number of decision trees GBMs do not necessarily employ all existing features in the data, but possibly only a subset. We found the optimal number of boosting steps found via an inner 10-fold cross-validation. Importantly, this was done within the outer 10 times repeated 10-fold cross-validation procedure used to evaluate prediction performance.

GBM can deal with censored time-to-event (here: time to AD diagnosis) data, as in our application: We like to predict the time until AD is diagnosed. For some AD converters such a diagnosis will be observed within the study time. However, there are also patients for which the diagnosis cannot be established within the study time, but potentially after the end of study. Their observed times to event are thus right censored. Our employed GBM implementation (R-package *gbm*) allows for dealing with time-to-event data by using the negative partial log-likelihood of the Cox proportional hazards model as a loss function.

As typical in clinical studies ADNI data contains missing values. GBM rely on a surrogate split approach for this purpose<sup>97</sup>. GBM allow for a ranking of variables according to their relevance for the model. This is done by recording the relative reduction of the error loss function as a measure of variable importance. Accordingly, features with zero importance can be filtered out. Hence, GBM can be used for feature selection.

In our data there is a difference in the number of features from SNPs, pathways, principal components and clinical data. In order to avoid any potential selection bias towards one of these feature types due to differing number of features we decided to implement a two-step strategy:

1. Training of a separate GBM model for each data modality and selection of most relevant features.
2. Joining of these features and training of a final GBM model.

Note that the first step ignores feature dependencies from different modalities while the second one takes them into account.

Multi-modal data fusion is a field of active research, and there is not a universal best performing approach. In the data science literature classically three general strategies for multi-modal data fusion are distinguished<sup>98,99</sup>, see Ahmad and Fröhlich<sup>100</sup> for a more extensive review<sup>100</sup>. Early data fusion methods focus on extraction of common features from several data modalities, resulting into one integrated data matrix. In a second step conventional machine learning methods can then be applied. Late integration algorithms first learn separate models for each data modality and then only combine predictions made by these models, for example with the help of a meta-model trained on the outputs of data source specific sub-models. These methods hence ignore feature dependencies between different data modalities. Intermediate integration algorithms are the youngest branch of data fusion approaches. The idea is to join data sources while building the predictive model. Our proposed approach can be seen as an instance of an intermediate data fusion approach.

No significant difference in prediction performance (C-index) compared to a conventional approach using a GBM model with all features was observed ( $p = 0.35$ , Wilcoxon test). However, the final GBM model with the two-step strategy contained fewer features (335 vs. 435).

**Comparison to Other Prediction Methods.** We compared our proposed approach to a Random Survival Forest<sup>15</sup>, resulting into a significantly higher C-index with our GBM modeling strategy ( $p = 1.1e-5$ , Wilcoxon test, Figure S4). The GBM model also outperformed an elastic net penalized Cox regression<sup>16,17</sup> ( $p = 0.0002$ , Wilcoxon test, Figure S4). Importantly, application of elastic net penalized Cox regression requires to impute missing values. This was done via the *missForest* algorithm<sup>101</sup> in a pre-processing step prior to running the cross-validation. Note that this step may result into slightly over-optimistic results for the elastic net.

In addition, we compared our approach against supervised sparse Generalized Canonical Correlation Analysis (ssGCCA)<sup>18,19</sup> in conjunction with a conventional Cox regression as predictive model. Sparse GCCA is an early data fusion approach that extracts latent variables (canonical variates) from different data modalities (here: clinical, SNPs, pathways, principal components). Each canonical variate describes a sparse linear combination of existing features within a specific data modality and is chosen to maximize the sum of correlations with canonical variates from other data modalities. We used a sparse GCCA version here, because we expect only a subset of original features to be relevant for the predicted outcome. In addition, we performed a supervised pre-filtering of features in each data modality. For this purpose and in agreement to Witten *et al.* we performed univariate Cox regressions and selected the 20% features with lowest p-value according to the log-likelihood ratio test. Afterwards we conducted feature extraction via sparse GCCA, as described before, and projected data of each modality into the low-dimensional space spanned by first canonical variates, and in that space a predictive Cox regression was trained. We here tested two different sparse GCCA implementations, one provided in R-package “*mixOmics*”<sup>102</sup> and the other one provided in R-package “*PMA*”<sup>18</sup>. Sparse GCCA involves tuning of regularization/sparsity parameters for each data modality. The sparse GCCA implementation in R-package “*PMA*” provides a permutation test for this purpose, while the sparse GCCA implementation in R-package “*mixOmics*” requires to run an inner cross-validation over a grid of regularization parameters (see details in Supplements). Both ssGCCA methods performed similar and resulted into significantly lower C-indices than our GBM approach (median 80% with *PMA*, 82% with *mixOmics* vs. 86% with our method;  $p < 1e-4$  for *PMA* and *mixOmics* vs. our method with

Wilcoxon test, see Figure S4). Omitting the supervised pre-filtering step suggested by Witten *et al.* resulted into a clear drop of the C-index by around 5% ( $p = 1.8e-5$  with Wilcoxon test, see Figure S4).

**Bayesian Network Learning.** *General.* Let  $G = (V, E)$  be a directed acyclic graph (DAG) and  $X = (X_v)_{v \in V}$  a set of random variables indexed by nodes in  $V$ .  $X$  is called a Bayesian Network (BN) with respect to  $G$ , if the joint distribution  $p(X_1, X_2, \dots, X_n)$  factorizes according to:

$$p(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = \prod_{v \in V} p(X_v = x_v | X_{pa(v)} = x_{pa(v)}) \quad (1)$$

where  $pa(v)$  denotes the parents of node  $v$  and  $x_{pa(v)}$  their joint configuration<sup>34</sup>.

The Markov Blanket (MB) of node  $v$ ,  $MB(v)$ , is defined as the set of nodes consisting of  $v$ 's parents, children, and any other parents of  $v$ 's children. If  $X$  is a BN with respect to  $G$ , then every node is conditionally independent of all other nodes in the network, given its Markov Blanket, i.e.

$$X_i \perp X_j | X_{MB(i)} \text{ for all } i \in V, j \in V - \{i\} - MB(i) \quad (2)$$

*Learning the Structure of a Bayesian Network from Data.* In the simplest case the DAG  $G$  is defined by an expert, but in many real life applications (as our present one) this is not the case and thus  $G$  should be learned from data. In general there exists two existing strategies for that purpose: search-and-score and constraint-based algorithms<sup>34</sup>. Search-and-score based approaches walk through the space of all possible DAG structures and score each candidate by its fit to the data. Typically such methods are thus computationally not scale-able to large BNs. In contrast, constrained-based approaches are significantly faster and scale-able to BNs with hundreds of variables. They typically rely on conditional independence tests between variables<sup>103</sup>. In this work we used and compared six different algorithms implemented in the R-package *bnlearn*<sup>104</sup>: greedy hill climbing (50 random restarts), tabu search<sup>38</sup>, Max-Min Hill Climbing (MMHC)<sup>105</sup>, Max-Min Parent Child (MMPC)<sup>105</sup> and semi-interleaved Hiton Parent Child (SI-HITON-PC)<sup>106</sup>. Greedy hill climbing and tabu search are heuristic score based optimization approaches, whereas MMPC and SI-HITON-PC are constrained-based structure learning methods that try to identify the Markov Blanket of each node in the Bayesian Network. MMHC is a hybrid approach, which uses ideas from both, search-and-score as well as constrained-based techniques: MMHC first learns the skeleton of the BN using the MMPC constrained-based algorithm. In a second phase edges are then oriented via a greedy hill climbing search.

Selection between different BN structure learning algorithms can be done via  $k$ -fold cross-validation akin to conventional supervised learning<sup>34</sup>. That means the overall data is randomly split into  $k$  folds, and the BN structure together with its parameters successively learned from  $k-1$  folds. If the fitted BN correctly models the overall population (and not just the training data), the data in the left out fold should with high probability fall into the same statistical distribution that is described by the BN. This can be quantified via the negated expected log-likelihood of the test data. Accordingly, cross-validation can be used to assess the generalization ability of a BN model and to compare different structure learning algorithms on that basis (see results in Supplements).

An important question in BN structure learning is, in how far the learned structure reflects causal relationships in the data. Indeed, if the BN is faithful to the underlying statistical distribution (i.e. models it correctly), then the true causal network is known to be part of a class of equivalent graph structures, called *class partially directed acyclic graph* (CPDAG)<sup>34,103</sup>. Under the above mentioned assumptions the CPDAG has the same skeleton as the true causal graph, but may leave some edges undirected. Hence, in practical applications it is important to restrict the CPDAG equivalence class as much as possible by prior knowledge to allow correct orientation of as many edges as possible. In our case we specifically imposed the following constraints for BN structures:

- No genomic feature can be influenced by a non-genomic feature. However, genomic features are allowed to have interactions among themselves (e.g. pathway-pathway dependencies).
- Neuro-psychological test results cannot be influenced by neuro-imaging features, but the other way around is possible.
- Age does not depend on any other variable.
- The baseline diagnosis cannot be influenced by any clinical variable, except for age and education level.
- The education level does not depend on any other clinical variable.
- The time-to-AD diagnosis is always dependent on a censoring indicator, and it does not influence any other variable.

Despite of these constraints identifying the true (CP)DAG structure from limited data still remains a challenge and thus raises the question, how confident one can be about the existence of an individual edge. One possible way of addressing this question is via a non-parametric bootstrap<sup>107</sup>. Briefly, given data from  $N$  patients we sample  $N$  patient records with replacement for a number of times (here: 1000). For each of these 1000 bootstrap samples a BN structure is learned. Afterwards the relative frequency of observing a particular edge is recorded, resulting into a confidence measure, which reflects the robustness of an edge against perturbations of the data.

Prior to BN structure learning we imputed missing values on the whole dataset of 900 patients using the miss-Forest algorithm for mixed categorical and continuous data types<sup>101</sup>. Furthermore, all variables were discretized into three bins using equal interval width.

**Comparison Against Literature Derived Cause-Effect-Relationships.** We refer to<sup>48</sup> for details about the construction of the literature derived, cause-and-effect relationship model for AD. One of the main challenges

with this model is that it contains many variables that have no direct correspondence in our data and vice versa. In our case only SNP rs405509 (in APOE4 gene) could be mapped directly to the OpenBEL model. To address this challenge we employed NeuroMMSig<sup>108</sup>. NeuroMMSig categorizes biological entities according to their role in disease specific pathways based on support from the literature. Corresponding references are stored in the NeuroMMSig database. This gene set view allowed us to relate OpenBEL sub-graphs to KEGG and Reactome pathways. For this purpose each KEGG and Reactome pathway was viewed as a gene set, using GeneCards<sup>109</sup> for gene to pathway mapping. For each pathway we then searched for the gene sets in the OpenBEL model with largest overlap. The statistical significance of this overlap was assessed via a hyper-geometric test and corrected for multiple testing using the Benjamini-Yekutieli method under dependency<sup>29</sup>. In conclusion 24 KEGG/Reactome pathways could be mapped at a significance cutoff of 5% (see complete list in Supplements).

Given two mapped gene sets  $A, B$  we then calculated shortest paths between all  $a \in A$  and  $b \in B$ . The union of these shortest paths was depicted as an OpenBEL sub-graph.

## References

- Burns, A. & Iliffe, S. Alzheimer's disease. *BMJ (Clinical research ed.)* **338**, b158 (2009).
- Jack, C. R. *et al.* Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *The Lancet. Neurol.* **9**, 119–128 (2010).
- Anoop, A., Singh, P. K., Jacob, R. S. & Maji, S. K. CSF Biomarkers for Alzheimer's Disease Diagnosis. *Int. J. Alzheimer's Dis.* **2010** (2010).
- Ferreira, D. *et al.* Meta-Review of CSF Core Biomarkers in Alzheimer's Disease: The State-of-the-Art after the New Revised Diagnostic Criteria. *Front. Aging Neurosci.* **6** (2014).
- Zhang, D., Wang, Y., Zhou, L., Yuan, H. & Shen, D. Multimodal Classification of Alzheimer's Disease and Mild Cognitive Impairment. *NeuroImage* **55**, 856–867 (2011).
- Cui, Y. *et al.* Identification of conversion from mild cognitive impairment to Alzheimer's disease using multivariate predictors. *PLoS One* **6**, e21896 (2011).
- Fan, Y., Batmanghelich, N., Clark, C. M. & Davatzikos, C. Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline. *NeuroImage* **39**, 1731–1743 (2008).
- Prestia, A. *et al.* Prediction of dementia in MCI patients based on core diagnostic markers for Alzheimer disease. *Neurol.* **80**, 1048–1056 (2013).
- Risacher, S. L. *et al.* Baseline MRI predictors of conversion from MCI to probable AD in the ADNI cohort. *Curr. Alzheimer Res.* **6**, 347–361 (2009).
- Lee, E. *et al.* Bflcrm: A Bayesian Functional Linear Cox Regression Model For Predicting Time To Conversion To Alzheimer's Disease. *The Annals Appl. Stat.* **9**, 2153–2178 (2015).
- Li, K., Chan, W., Doody, R. S., Quinn, J. & Luo, S. Prediction of conversion to Alzheimer's disease with longitudinal measures and time-to-event data. *J. Alzheimer's disease: JAD* **58**, 361–371 (2017).
- Friedman, J. H. Stochastic gradient boosting. *Comput. Stat. & Data Analysis* **38**, 367–378 (2002).
- Mehenni, T. & Moussaoui, A. Data mining from multiple heterogeneous relational databases using decision tree classification. *Pattern Recognit. Lett.* **33**, 1768–1775 (2012).
- Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L. & Rosati, R. A. Evaluating the yield of medical tests. *JAMA* **247**, 2543–2546 (1982).
- Ishwaran, H., Kogalur, U. B., Blackstone, E. H. & Lauer, M. S. Random survival forests. *The Annals Appl. Stat.* **2**, 841–860 (2008).
- Zou, H. & Hastie, T. Regularization and variable selection via the Elastic Net. *J. Royal Stat. Soc. Ser. B* **67**, 301–320 (2005).
- Wu, Y. Elastic Net For Cox's Proportional Hazards Model With A Solution Path Algorithm. *Stat. Sinica* **22**, 27–294 (2012).
- Witten, D. M. & Tibshirani, R. J. Extensions of sparse canonical correlation analysis with applications to genomic data. *Stat. applications in genetics and molecular biology* **8**, 1–27 (2009).
- Tenenhaus, A. *et al.* Variable selection for generalized canonical correlation analysis. *Biostat. (Oxford, England)* **15**, 569–583 (2014).
- Moh, C. *et al.* Cell cycle deregulation in the neurons of Alzheimer's disease. *Results Probl. Cell Differ.* **53**, 565–576 (2011).
- Nagy, Z., Esiri, M. M. & Smith, A. D. The cell division cycle and the pathophysiology of Alzheimer's disease. *Neurosci.* **87**, 731–739 (1998).
- Sadigh-Eteghad, S. *et al.* Amyloid-Beta: A Crucial Factor in Alzheimer's Disease. *Med. Princ. Pract.* **24**, 1–10 (2015).
- Heneka, M. T., Golenbock, D. T. & Latz, E. Innate immunity in Alzheimer's disease. *Nat. Immunol.* **16**, 229–236 (2015).
- Ding, Q., Markesbery, W. R., Chen, Q., Li, F. & Keller, J. N. Ribosome dysfunction is an early event in Alzheimer's disease. *The J. Neurosci. The Off. J. Soc. for Neurosci.* **25**, 9171–9175 (2005).
- Furney, S. J. *et al.* Genome-wide association with MRI atrophy measures as a quantitative trait locus for Alzheimer's disease. *Mol. Psychiatry* **16**, 1130–1138 (2011).
- Bufill, E. *et al.* Reelin signaling pathway genotypes and Alzheimer disease in a Spanish population. *Alzheimer Dis. Assoc. Disord.* **29**, 169–172 (2015).
- Chouraki, V. *et al.* A genome-wide association meta-analysis of plasma A $\beta$  peptides concentrations in the elderly. *Mol. psychiatry* **19**, 1326–1335 (2014).
- Abraham, R. *et al.* A genome-wide association study for late-onset Alzheimer's disease using DNA pooling. *BMC medical genomics* **1**, 44 (2008).
- Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Annals Stat.* **29**, 1165–1188 (2001).
- Sherry, S. T. *et al.* dbSNP: The NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
- Ma, C. *et al.* The TT allele of rs405509 synergizes with APOE  $\epsilon$ 4 in the impairment of cognition and its underlying default mode network in non-demented elderly. *Curr. Alzheimer Res.* **13**, 708–717 (2016).
- Zou, Y.-m, Lu, D., Liu, L.-p, Zhang, H.-h & Zhou, Y.-y Olfactory dysfunction in Alzheimer's disease. *Neuropsychiatr. Dis. Treat.* **12**, 869–875 (2016).
- Gondi, C. S., Dinh, D. H., Klopfenstein, J. D., Gujrati, M. & Rao, J. S. MMP-2 Downregulation Mediates Differential Regulation of Cell Death via ErbB-2 in Glioma Xenografts. *Int. journal oncology* **35**, 257–263 (2009).
- Koller, D. & Friedman, N. *Probabilistic Graphical Models: Principles and Technique* (MIT Press, 2009).
- Sachs, K., Perez, O., Peèr, D., Lauffenburger, D. & Nolan, G. Causal protein-signaling networks derived from multiparameter single-cell data. *Sci.* **208**, 523–529 (2005).
- Friedman, N. Inferring cellular networks using probabilistic graphical models. *Sci.* **303**, 799–805 (2004).
- Friedman, N., Linial, M., Nachman, I. & Peèr, D. Using Bayesian networks to analyze expression data. *J. Comput. Biol* **7**, 601–620 (2000).
- Hong, Y., Xia, X., Le, J. & Zhou, X. Learning Bayesian Network Structure from Large-Scale Datasets. In *2016 International Conference on Advanced Cloud and Big Data (CBD)*, 258–264 (2016).

39. Van Hoesen, G. W., Hyman, B. T. & Damasio, A. R. Entorhinal cortex pathology in Alzheimer's disease. *Hippocampus* **1**, 1–8 (1991).
40. Liu, C.-C., Kanekiyo, T., Xu, H. & Bu, G. Apolipoprotein E and Alzheimer disease: Risk, mechanisms, and therapy. *Nat. reviews. Neurol.* **9**, 106–118 (2013).
41. Mu, Y. & Gage, F. H. Adult hippocampal neurogenesis and its role in Alzheimer's disease. *Mol. Neurodegener.* **6**, 85 (2011).
42. Coppède, F. & Migliore, L. DNA damage and repair in Alzheimer's disease. *Curr. Alzheimer Res.* **6**, 36–47 (2009).
43. Shiwaku, H. & Okazawa, H. Impaired DNA damage repair as a common feature of neurodegenerative diseases and psychiatric disorders. *Curr. Mol. Medicine* **15**, 119–128 (2015).
44. Nava-Mesa, M. O., Jiménez-Daz, L., Yajeya, J. & Navarro-Lopez, J. D. GABAergic neurotransmission and new strategies of neuromodulation to compensate synaptic dysfunction in early stages of Alzheimer's disease. *Front. Cell. Neurosci.* **8** (2014).
45. Koppel, J. & Davies, P. Targeting the Endocannabinoid System in Alzheimer's Disease. *J. Alzheimer's disease: JAD* **15**, 495–504 (2008).
46. Kanehisa, M. *et al.* KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* **36**, 480–484 (2008).
47. Fabregat, A. *et al.* The Reactome pathway Knowledgebase. *Nucleic Acids Res.* **44**, D481–487 (2016).
48. Kodamullil, A. T., Younesi, E., Naz, M., Bagewadi, S. & Hofmann-Apitius, M. Computable cause-and-effect models of healthy and Alzheimer's disease states and their mechanistic differential analysis. *Alzheimer's & Dementia* **11**, 1329–1339 (2015).
49. Nighot, P. & Ma, T. Role of autophagy in the regulation of epithelial cell junctions. *Tissue Barriers* **4** (2016).
50. Tietz, S. & Engelhardt, B. Brain barriers: Crosstalk between complex tight junctions and adherens junctions. *J. Cell Biol.* **209**, 493–506 (2015).
51. Stamatovic, S. M., Keep, R. F. & Andjelkovic, A. V. Brain endothelial cell-cell junctions: how to “open” the blood brain barrier. *Curr. neuropharmacology* **6**, 179–92 (2008).
52. Weiss, N., Miller, F., Cazaubon, S. & Couraud, P. O. The blood-brain barrier in brain homeostasis and neurological diseases. *Biochimica et Biophys. Acta - Biomembr.* **1788**, 842–857 (2009).
53. Zenaro, E., Piacentino, G. & Constantin, G. The blood-brain barrier in Alzheimer's disease. *Neurobiol. Dis.* (2016).
54. Alvira-Botero, X. *et al.* Megalin interacts with APP and the intracellular adapter protein FE65 in neurons. *Mol. Cell. Neurosci.* **45**, 306–315 (2010).
55. Jiang, S. *et al.* Trafficking regulation of proteins in Alzheimer's disease. *Mol. Neurodegener.* **9**, 6 (2014).
56. Zhang, X. & Song, W. The role of APP and BACE1 trafficking in APP processing and amyloid- $\beta$  generation. *Alzheimer's research & therapy* **5**, 46 (2013).
57. Lee, E., Chuang, H.-Y., Kim, J.-W., Ideker, T. & Lee, D. Inferring pathway activity toward precise disease classification. *PLoS Comput. Biol.* **4**, e1000217 (2008).
58. Tamayev, R., Zhou, D. & D'Adamo, L. The interactome of the amyloid beta precursor protein family members is shaped by phosphorylation of their intracellular domains. *Mol. Neurodegener.* **4**, 28 (2009).
59. Shrivastava-Ranjan P, Faundez, V. & Fang, G. *et al.* Int3/X11 $\gamma$  Is an ADP-Ribosylation Factor-dependent Adaptor that Regulates the Traffic of the Alzheimer's Precursor Protein from the Trans-Golgi Network. *Mol. Biol. Cell.* (2008).
60. King, G. D., Perez, R. G., Steinhilb, M. L., Gaut, T. R., JR X11 $\alpha$  modulates secretory and endocytic trafficking and metabolism of amyloid precursor protein: mutational analysis of the YENPTY sequence. *Neurosci.* (2003).
61. Clarke, J. L. & Daniell, H. Plastid biotechnology for crop production: Present status and future perspectives. *Plant Mol. Biol.* **76**, 211–220 (2011).
62. Montesperan, C., Wiethoff, C. M. & Wodrich, H. A small viral PPxY-peptide motif to control antiviral autophagy. *J. of Virol.* JVI.00581–17 (2017).
63. Funderburk, S., Marcellino, B. & Yue, Z. Cell “Self Eating” (Autophagy) Mechanism in Alzheimer's Disease. *Mt. Sinai J. Medicine* **77**, 59–68 (2010).
64. Fu, M. M. & Holzbaur, E. L. F. MAPK8IP1/JIP1 regulates the trafficking of autophagosomes in neurons. *Autophagy* **10**, 2079–2081 (2014).
65. Ariosa, A. R. & Klionsky, D. J. Autophagy core machinery: overcoming spatial barriers in neurons. *J. Mol. Medicine* **94**, 1217–1227 (2016).
66. Chauhan, S. *et al.* Pharmaceutical screen identifies novel target processes for activation of autophagy with a broad translational potential. *Nat. communications* **6**, 8620 (2015).
67. Pickford, F. *et al.* The autophagy-related protein beclin 1 shows reduced expression in early Alzheimer disease and regulates amyloid  $\beta$  accumulation in mice. *J. Clin. Investig.* **118**, 2190–2199 (2008).
68. Lorini, R. *et al.* Cytotoxic activity in children with insulin-dependent diabetes mellitus. *Diabetes Res. Clin. Pract.* **23**, 37–42 (1994).
69. Maday, S. *et al.* Axonal Transport: Cargo-Specific Mechanisms of Motility and Regulation. *Neuron* (2014).
70. Muresan, V. & Muresan, Z. A persistent stress response to impeded axonal transport leads to accumulation of amyloid- $\beta$  in the endoplasmic reticulum, and is a probable cause of sporadic Alzheimer's disease. *Neurodegener. Dis.* **10**, 60–63 (2012).
71. Szodorai, A. *et al.* APP Anterograde Transport Requires Rab3A GTPase Activity for Assembly of the Transport Vesicle. *J. Neurosci.* **29**, 14534–14544 (2009).
72. Craft, S. & Watson, G. S. Insulin and neurodegenerative disease: shared and specific mechanisms. *Lancet Neurol.* (2004).
73. Hoyer, S. Glucose metabolism and insulin receptor signal transduction in Alzheimer disease. *Eur. J. Pharmacol.* (2004).
74. Zhao, W. Q., Chen, H., Quon, M. J. & Alkon, D. L. Insulin and the insulin receptor in experimental models of learning and memory. *Eur. J. Pharmacol.* (2004).
75. Banks, W. A., Owen, J. B. & Erickson, M. A. Insulin in the brain: There and back again. *Pharmacol. Ther.* **136**, 82–93 (2012).
76. Duarte, A. I., Moreira, P. I. & Oliveira, C. R. Insulin in central nervous system: More than just a peripheral hormone. *J. Aging Res.* **2012** (2012).
77. Blázquez, E., Velázquez, E., Hurtado-Carneiro, V. & Ruiz-Albusac, J. M. Insulin in the Brain: Its Pathophysiological Implications for States Related with Central Insulin Resistance, Type 2 Diabetes and Alzheimer's Disease. *Frontiers in Endocrinology* **5** (2014).
78. Conejo, R., Lorenzo, M. Insulin signaling leading to proliferation, survival, and membrane ruffling in C2C12 myoblasts. *J. cellular physiology* (2001).
79. Bifulco, G., *et al.* Glucose regulates insulin mitogenic effect by modulating SHP-2 activation and localization in Jar cells. *The journal of biological chemistry* (2002).
80. Yarchoan, M. *et al.* Abnormal serine phosphorylation of insulin receptor substrate 1 is associated with tau pathology in Alzheimer's disease and tauopathies. *Acta Neuropathol.* **128**, 679–689 (2014).
81. Coppers, K. D. & White, M. F. Regulation of insulin sensitivity by serine/threonine phosphorylation of insulin receptor substrate proteins IRS1 and IRS2. *Diabetologia* (2012).
82. de la Monte, S. M. & Wands, J. R. Alzheimer's Disease Is Type 3 Diabetes - Evidence Reviewed. *J. Diabetes Sci. Technol.* **2**, 1101–1113 (2008).
83. Talbot, K. *et al.* Demonstrated brain insulin resistance in Alzheimer's disease patients is associated with IGF-1 resistance, IRS-1 dysregulation, and cognitive decline. *J. Clin. Investig.* **122**, 1316–1338 (2012).
84. Holtzman, D. M., John, C. M. & Goate, A. Alzheimer's Disease: The Challenge of the Second Century. *Sc. translational medicine* **3**, 77sr1 (2011).

85. Wheeler, D. L. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **35**, D5–D12 (2007).
86. Purcell, S. *et al.* PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
87. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
88. Zheng, X. *et al.* A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinforma. (Oxford, England)* **28**, 3326–3328 (2012).
89. Younesi, E. *et al.* Mining biomarker information in biomedical literature. *BMC Med. Informatics Decis. Mak.* **12**, 148 (2012).
90. Piñero, J. *et al.* DisGeNET: A comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.* **45**, D833–D839 (2017).
91. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2017).
92. Nounou, M. N., Bakshi, B. R., Goel, P. K. & Shen, X. Bayesian principal component analysis. *J. Chemom.* **16**, 576–595 (2002).
93. Lanczos, C. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *J. Res. Natl. Bur. Stand. B* **45**, 255–282 (1950).
94. Carithers, L. J. *et al.* A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. *Biopreservation and Biobanking* **13**, 311–319 (2015).
95. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4**, 1073–1081 (2009).
96. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. *Curr. protocols human genetics* **07**, Unit7.20 (2013).
97. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning*, 2 edn (Springer, New York, NY, USA, 2008).
98. Pavlidis, P., Weston, J., Cai, J. & Grundy, W. N. Gene Functional Classification from Heterogeneous Data. In *Proceedings of the Fifth Annual International Conference on Computational Biology*, RECOMB '01, 249–255 (ACM, New York, NY, USA, 2001).
99. Maragos, P., Gros, P., Katsamanis, A. & Papandreou, G. Cross-Modal Integration for Performance Improving in Multimedia: A Review. In Maragos, P., Potamianos, A. & Gros, P. (eds.) *Multimodal Processing and Interaction*, 1–46 (Springer US, Boston, MA, 2008).
100. Ahmad, A. & Fröhlich, H. Integrating Heterogeneous omics Data via Statistical Inference and Learning Techniques. *Genomics Comput. Biol.* **2**, 32 (2016).
101. Stekhoven, D. J. & Bühlmann, P. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinforma. (Oxford, England)* **28**, 112–118 (2012).
102. Rohart, F., Gautier, B., Singh, A. & Cao, K.-A. L. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLOS Computat. Biol.* **13**, e1005752 (2017).
103. Spirtes, P., Glymour, C. N. & Scheines, R. *Causation, Prediction, and Search*, vol. 81 (MIT press, 2000).
104. Scutari, M. Learning Bayesian Networks with the bnlearn R Package. *J. Stat. Softw.* **35**, 1–22 (2010).
105. Tsamardinos, I., Brown, L. & Aliferis, C. The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm. *Machine Learning* **65**, 31–78 (2006).
106. Aliferis, C. F., Statnikov, A., Tsamardinos, I. & Mani, S. & Koutsoukos, X. Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification Part I: Algorithms and Empirical Evaluation. *J. Mach. Learn. Res.* **11**, 171–234 (2010).
107. Friedman, N., Goldszmidt, M. & Wyner, A. Data Analysis with Bayesian Networks: A Bootstrap Approach. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, UAI'99, 196–205 (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1999).
108. Domingo-Fernandez, D. *et al.* Multimodal Mechanistic Signatures for Neurodegenerative Diseases (NeuroMMSig): A web server for mechanism enrichment. *Bioinforma.* (2017).
109. Safran, M. *et al.* GeneCards Version 3: The human gene integrator. *Database: The J. Biol. Databases Curation* **2010**, baq020 (2010).

## Acknowledgements

This project has been partially supported by the IMI project AETIONOMY (<https://www.aetionomy.eu/en/vision.html>) within 7th framework program of the European Union.

## Author Contributions

S.K. implemented and tested machine learning models and performed genomic data analysis, D.D.-F. implemented the mapping of BN features to OpenBEL graphs, A.I. helped with the biological interpretation of results, M.A.E. helped with genomic data analysis, H.F. and M.H.-A. designed the research, H.F. guided the project and drafted the manuscript. All authors reviewed and approved the manuscript.

## Additional Information

**Supplementary information** accompanies this paper at <https://doi.org/10.1038/s41598-018-29433-3>.

**Competing Interests:** The authors declare no competing interests.

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



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

© The Author(s) 2018



## Conclusions

The work presents a novel strategy to bridge the gap between clinical data and mechanistic knowledge that not only improves the robustness of a predictive model but also its interpretability. We demonstrate that the most predictive multi-scale features from our model can be used to reconstruct the biological mechanisms that are known to play a role in AD pathophysiology. As future work, the pinpointed mechanisms could be targeted and further validated in a clinical study. Further, the generalizability and success of this approach has made it de facto a state-of-the-art methodology as it is currently employed on other domains such as PD, more specifically on the PPMI dataset.





# 8

## Conclusion and outlook

This new age of information, where data and computational power are constantly increasing [130], has shifted numerous paradigms in the biomedical field. While in the past, the challenge was to find patterns in data, the present asks for novel methods devised to assist in interpreting patterns emerging from the vast amount of data. One of the reasons explaining this phenomena is the increasing mismatch between the amount of signals coming from *-omics* experiments and the lack of means to filter and decode these signals. Therefore, enhancing interpretation power is crucial, especially in interdisciplinary fields such as bioinformatics.

This dissertation has presented a repertoire of both knowledge- and data-driven computational methods designed to model complex biological systems. First, this work has shown novel knowledge-based methods that leverage pathway and mechanistic knowledge to drive interpretation and analysis of biomedical data. Additionally, this thesis has presented sophisticated machine learning methods that illustrate how both approaches (i.e., knowledge- and data-driven) can synergistically complement each other, in particular, in the challenging area of neurology. Besides the methodological aspect of this work, the integrative efforts conducted provide a more comprehensive view of the canonical pathway landscape. Further, this work has presented the world's largest mechanistic inventory in the neurodegeneration area containing over two hundred highly-curated and context specific networks modeling disease mechanisms. This immense knowledge template can be used for patient stratification based on disease mechanisms; thereby, improving diagnosis and personalizing treatments. In summary, the mechanistic knowledge contained in the pathway networks and disease maps presented in this thesis have the potential to widen the knowledge space by revealing novel interactions and crosstalks that could uncover the pathophysiology underlying human conditions, particularly in the neurodegeneration and psychiatric arena.

This thesis has added to the bioinformatics field by establishing the first pathway mappings and harmonizing pathway representations across major databases. By overcoming the two major obstacles in pathway knowledge integration (i.e., linking pathway concepts and unifying formats), this work has described a method to integrate knowledge from multiple resources. To conduct this endeavor, two ecosystems (i.e., ComPath and PathMe) were implemented using state of the art technologies and following a modular design to facilitate their re-usability. Therefore, both open source tools can be used to keep track of rapid changes in the pathway landscape as well as to incorporate additional resources in the future. Finally, another important aspect of the work part of this thesis is its strong focus on reproducibility. All the resources, software, analytical scripts or pipelines presented here have been packaged and tested following strict software development standards by the scientific community [131, 132] and are all publicly available in major repositories such as GitHub or PyPi.

Beyond the direct link with pathway analyses, both resources are part of the Bio2BEL framework [41], so they can be used for other applications such as the enrichment of BEL networks. As an illustration, the ComPath mapping dataset has been used to demonstrate the promising application of network representation learning approaches in the biomedical domain [133]. Furthermore, the integration and consolidation of three major pathway resources is instrumental in generating novel datasets that exploit the benefits of multiple resources and facilitate the dissemination of pathway knowledge. This is strongly evidenced by the benchmarking study presented in this thesis which also introduces an integrative approach that paves the way for a future consolidated pathway landscape.

Our benchmarking study not only shows how instrumental integrative approaches are in order to reach consensus conclusions, but also raises awareness about the negative impact of restricting analysis to individual databases. This, in turn, opens debate about the reproducibility of pathway-driven analyses and strongly encourages the scientific community to adopt integrative approaches aimed at unifying pathway knowledge in order to mitigate this bias. Furthermore, the software implemented during the course of the study, PathwayForte, enables a re-evaluation of the presented results as well as the incorporation of novel databases and methods in the future. Finally, high-impact scientific publications employing pathway-centric methods could be critically evaluated based on database selection. Such analysis would illustrate the potential issue of resource *cherry picking* in our field (i.e., running analyses on different databases until you are satisfied with results).

In principle, the pathway-centric approaches presented in this dissertation can

be applied to any disease domain with little or no further adaptations. Pathway databases drive mechanistic interpretation of disease aetiology in areas where both data and prior knowledge are abundant. However, complex and poorly-understood indications require tailored approaches where knowledge is contextualized and is exclusively derived from the particular disease or molecular process of interest. As an example, after decades of research in AD and PD, little is known about the molecular basis underlying these disorders. Therefore, one of the goals of this thesis has been to organize mechanistic knowledge in the neurodegenerative arena by curating and modeling disease-specific mechanisms by leveraging previous work [134].

In developing NeuroMMSig, we have established the first draft of a mechanism-based taxonomy for AD and PD, constituting a central part of the AETIONOMY project<sup>1</sup>. On-going work in this project focuses on using the definitions of pathways for clustering dementia patients. We have also demonstrated how researchers are able to investigate shared mechanisms across diseases thanks to the common schemata and semantic alignment followed through the mechanism inventory [123]. The context of NeuroMMSig has been extended with new diseases such as epilepsy [123]. Further, NeuroMMSig was extended during the Human Brain Pharmacome project<sup>2</sup> with chemogenomic information to support computational prediction of drug repositioning candidates. This work has also been used to support data-driven analysis of differential gene expression profiles of AD using heat diffusion algorithms [136]. Thus, future efforts could be directed towards enabling the submission of multimodal clinical data as outlined by [135], or using novel algorithms on the mechanism enrichment server. Ultimately, NeuroMMSig has been a teaching tool that lead clinicians, wet-lab scientists, and bioinformaticians to their first steps into the world of systems and networks biology. In the future, further work can be conducted to transfer the paradigm of NeuroMMSig to other disease domains such as the psychiatric arena. In summary, the vast collection of mechanistic networks is not only instrumental for identifying the mechanisms underlying disparate subtypes of these conditions, but it is also a technology enabler that can have future implications in designing targets for other mechanisms rather than the ones the pharmaceutical industry have been focused on in the last decades [87].

Another endeavor of this thesis focused on cataloging biomarker knowledge in the area of PTSD by developing the first biomarker database for this disorder: PTSDDB. This functional resource includes a wide spectrum of biomarkers an-

---

<sup>1</sup><https://www.aetionomy.eu>

<sup>2</sup><https://www.pharmacome.scai.fraunhofer.de>

alyzed for this indication from over a hundred different studies and provides a suite of visualizations that enable the exploration of its highly-curated content. Further, this work demonstrates how displaying biomarker knowledge through user-friendly and interactive tools not only reveals new insights by examining what has already been achieved in the biomarker landscape, but also guides the design of future clinical, statistical, and bioinformatics studies. As an example of one of its multiple applications, we showed how this database can easily identify conflicting literature in the field. Ultimately, the highly-curated content incorporated in PTSDDDB can be used to conduct a meta-analysis study around certain biomarkers of interest.

Apart from pathway enrichment methods, few approaches that synergically leverage data- and knowledge-driven approaches for data interpretation exist. However, validating hypotheses derived from either approach is essential to prioritize candidates in the first steps of the drug development process. The final publication of this thesis has demonstrated how it is possible to enhance the interpretation in translational clinical research. Our approach first reveals the dependencies across the most predictive features via Bayesian modeling and linking them to knowledge-derived networks (i.e., NeuroMMSig mechanisms). Since this methodology was applied in the AD ADNI clinical cohort, the promising mechanistic links that were pinpointed could be further validated in an independent study. Looking forward, this novel approach also paves the way to include time dimension and thus disease progression into pathway and mechanistic networks. By doing so, we would be able to link mechanistic hypotheses to progression models and reveal how mechanisms dynamically change over time.

# Bibliography

- [1] H. Fröhlich, R. Balling, N. Beerenwinkel, O. Kohlbacher, S. Kumar, T. Lengauer, M. H. Maathuis, Y. Moreau, S. A. Murphy, T. M. Przytycka, M. Rebhan, H. Röst, A. Schuppert, M. Schwab, R. Spang, D. Stekhoven, J. Sun, A. Weber, D. Ziemek, B. Zupan. From hype to reality: data science enabling personalized medicine. *BMC medicine* 1 **Aug. 2018**, 16, 150.
- [2] F. Castiglione, F. Pappalardo, C. Bianca, G. Russo, S. Motta. Modeling biology spanning different scales: an open challenge. *BioMed research international* **2014**, 2014, 902545.
- [3] Z. Qu, A. Garfinkel, J. N. Weiss, M. Nivala. Multi-scale modeling in biology: how to bridge the gaps between scales? *Progress in biophysics and molecular biology* 1 **Oct. 2011**, 107, 21–31.
- [4] J. Walpole, J. A. Papin, S. M. Peirce. Multiscale computational models of complex biological systems. *Annual review of biomedical engineering* **2013**, 15, 137–154.
- [5] F. Uschner, E. Klipp. Signaling pathways in context. *Current opinion in biotechnology* **Apr. 2019**, 58, 155–160.
- [6] M. P. Cary, G. D. Bader, C. Sander. Pathway information for systems biology. *FEBS letters* 8 **Mar. 2005**, 579, 1815–1820.
- [7] D. Domingo-Fernández, C. T. Hoyt, C. Bobis-Álvarez, J. Marín-Llaó, M. Hofmann-Apitius. ComPath: an ecosystem for exploring, analyzing, and curating mappings across pathway databases. *NPJ systems biology and applications* **2019**, 5, 3.
- [8] M. Gündel, C. T. Hoyt, M. Hofmann-Apitius. BEL2ABM: agent-based simulation of static models in Biological Expression Language. *Bioinformatics* 13 **July 2018**, 34, 2316–2318.
- [9] F. Llaneras, J. Picó. A procedure for the estimation over time of metabolic fluxes in scenarios where measurements are uncertain and/or insufficient. *BMC bioinformatics* **Oct. 2007**, 8, 421.

- [10] Z. P. Gerdtzen. Modeling metabolic networks for mammalian cell systems: general considerations, modeling strategies, and available tools. *Advances in biochemical engineering/biotechnology* **2012**, 127, 71–108.
- [11] J. Hou, L. Acharya, D. Zhu, J. Cheng. An overview of bioinformatics methods for modeling biological pathways in yeast. *Briefings in functional genomics* **2 Mar. 2016**, 15, 95–108.
- [12] C. J. Needham, J. R. Bradford, A. J. Bulpitt, D. R. Westhead. A primer on learning in Bayesian networks for computational biology. *PLoS computational biology* **8 Aug. 2007**, 3, e129.
- [13] C. T. Cummings, D. Deryckere, H. S. Earp, D. K. Graham. Molecular pathways: MERTK signaling in cancer. *Clinical cancer research: an official journal of the American Association for Cancer Research* **19 Oct. 2013**, 19, 5275–5280.
- [14] J. Reimand, R. Isserlin, V. Voisin, M. Kucera, C. Tannus-Lopes, A. Rostamianfar, L. Wadi, M. Meyer, J. Wong, C. Xu, D. Merico, G. D. Bader. Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nature protocols* **2 Feb. 2019**, 14, 482–517.
- [15] G. Bader. *Pathguide: the pathway resource list*. URL: <http://pathguide.org/> (visited on 08/08/2019).
- [16] G. D. Bader, M. P. Cary, C. Sander. Pathguide: a pathway resource list. *Nucleic acids research Database issue* **Jan. 2006**, 34, D504–D506.
- [17] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, K. Morishima. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic acids research D1* **Jan. 2017**, 45, D353–D361.
- [18] A. Fabregat, K. Sidiropoulos, P. Garapati, M. Gillespie, K. Hausmann, R. Haw, B. Jassal, S. Jupe, F. Korninger, S. McKay, L. Matthews, B. May, M. Milacic, K. Rothfels, V. Shamovsky, M. Webber, J. Weiser, M. Williams, G. Wu, L. Stein, H. Hermjakob, P. D'Eustachio. The Reactome pathway Knowledgebase. *eng. Nucleic Acids Res* **Jan. 2016**, 44 (D1), D481–D487.
- [19] D. Croft, A. F. Mundo, R. Haw, M. Milacic, J. Weiser, G. Wu, M. Caudy, P. Garapati, M. Gillespie, M. R. Kamdar, B. Jassal, S. Jupe, L. Matthews, B. May, S. Palatnik, K. Rothfels, V. Shamovsky, H. Song, M. Williams, E. Birney, H. Hermjakob, L. Stein, P. D'Eustachio. The Reactome pathway knowledgebase. *eng. Nucleic Acids Res* **Jan. 2014**, 42 (Database issue), D472–D477.

- [20] D. N. Slenter, M. Kutmon, K. Hanspers, A. Riutta, J. Windsor, N. Nunes, J. Mélius, E. Cirillo, S. L. Coort, D. Digles, F. Ehrhart, P. Giesbertz, M. Kalafati, M. Martens, R. Miller, K. Nishida, L. Rieswijk, A. Waagmeester, L. M. T. Eijssen, C. T. Evelo, A. R. Pico, E. L. Willighagen. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic acids research* D1 **Jan. 2018**, 46, D661–D667.
- [21] A. R. Pico, T. Kelder, M. P. van Iersel, K. Hanspers, B. R. Conklin, C. Evelo. WikiPathways: pathway editing for the people. *PLoS biology* 7 **July 2008**, 6, e184.
- [22] M. Kutmon, A. Riutta, N. Nunes, K. Hanspers, E. L. Willighagen, A. Bohler, J. Mélius, A. Waagmeester, S. R. Sinha, R. Miller, S. L. Coort, E. Cirillo, B. Smeets, C. T. Evelo, A. R. Pico. WikiPathways: capturing the full diversity of pathway knowledge. *Nucleic acids research* D1 **Jan. 2016**, 44, D488–D494.
- [23] A. Bohler, G. Wu, M. Kutmon, L. A. Pradhana, S. L. Coort, K. Hanspers, R. Haw, A. R. Pico, C. T. Evelo. Reactome from a WikiPathways Perspective. *PLoS computational biology* 5 **May 2016**, 12, e1004941.
- [24] H. Mi, P. Thomas. PANTHER pathway: an ontology-based pathway database coupled with data analysis tools. *Methods in molecular biology (Clifton, N.J.)* **2009**, 563, 123–140.
- [25] E. G. Cerami, B. E. Gross, E. Demir, I. Rodchenkov, O. Babur, N. Anwar, N. Schultz, G. D. Bader, C. Sander. Pathway Commons, a web resource for biological pathway data. *Nucleic acids research* Database issue **Jan. 2011**, 39, D685–D690.
- [26] A. Kamburov, U. Stelzl, H. Lehrach, R. Herwig. The ConsensusPathDB interaction database: 2013 update. *Nucleic acids research* Database issue **Jan. 2013**, 41, D793–D800.
- [27] A. Kamburov, K. Pentchev, H. Galicka, C. Wierling, H. Lehrach, R. Herwig. ConsensusPathDB: toward a more complete picture of cell biology. *Nucleic acids research* Database issue **Jan. 2011**, 39, D712–D717.
- [28] V. Petri, P. Jayaraman, M. Tutaj, G. T. Hayman, J. R. Smith, J. De Pons, S. J. Laudederkind, T. F. Lowry, R. Nigam, S.-J. Wang, M. Shimoyama, M. R. Dwinell, D. H. Munzenmaier, E. A. Worthey, H. J. Jacob. The pathway ontology - updates and applications. *Journal of biomedical semantics* 1 **Feb. 2014**, 5, 7.
- [29] M. L. Green, P. D. Karp. The outcomes of pathway database computations depend on pathway ontology. *Nucleic acids research* 13 **2006**, 34, 3687–3697.

- [30] F. Belleau, M.-A. Nolin, N. Tourigny, P. Rigault, J. Morissette. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *Journal of biomedical informatics* 5 Oct. 2008, 41, 706–716.
- [31] A. Waagmeester, M. Kutmon, A. Riutta, R. Miller, E. L. Willighagen, C. T. Evelo, A. R. Pico. Using the Semantic Web for Rapid Integration of WikiPathways with Other Biological Online Data Resources. *PLoS computational biology* 6 June 2016, 12, e1004989.
- [32] F. Å. Nielsen, D. Mietchen, E. Willighagen. Scholia, Scientometrics and Wikidata. In: *The Semantic Web: ESWC 2017 Satellite Events*. Cham: Springer International Publishing, 2017, 237–259.
- [33] E. Demir, M. P. Cary, S. Paley, K. Fukuda, C. Lemer, I. Vastrik, G. Wu, P. D’Eustachio, C. Schaefer, J. Luciano, F. Schacherer, I. Martinez-Flores, Z. Hu, V. Jimenez-Jacinto, G. Joshi-Tope, K. Kandasamy, A. C. Lopez-Fuentes, H. Mi, E. Pichler, I. Rodchenkov, A. Splendiani, S. Tkachev, J. Zucker, G. Gopinath, H. Rajasimha, R. Ramakrishnan, I. Shah, M. Syed, N. Anwar, O. Babur, M. Blinov, E. Brauner, D. Corwin, S. Donaldson, F. Gibbons, R. Goldberg, P. Hornbeck, A. Luna, P. Murray-Rust, E. Neumann, O. Ruebenacker, O. Reubenacker, M. Samwald, M. van Iersel, S. Wimalaratne, K. Allen, B. Braun, M. Whirl-Carrillo, K.-H. Cheung, K. Dahlquist, A. Finney, M. Gillespie, E. Glass, L. Gong, R. Haw, M. Honig, O. Hubaut, D. Kane, S. Krupa, M. Kutmon, J. Leonard, D. Marks, D. Merberg, V. Petri, A. Pico, D. Ravenscroft, L. Ren, N. Shah, M. Sunshine, R. Tang, R. Whaley, S. Letovksy, K. H. Buetow, A. Rzhetsky, V. Schachter, B. S. Sobral, U. Dogrusoz, S. McWeeney, M. Aladjem, E. Birney, J. Collado-Vides, S. Goto, M. Hucka, N. Le Novère, N. Maltsev, A. Pandey, P. Thomas, E. Wingender, P. D. Karp, C. Sander, G. D. Bader. The BioPAX community standard for pathway data sharing. *Nature biotechnology* 9 Sept. 2010, 28, 935–942.
- [34] M. Hucka, A. Finney, H. M. Sauro, H. Bolouri, J. C. Doyle, H. Kitano, A. P. Arkin, B. J. Bornstein, D. Bray, A. Cornish-Bowden, A. A. Cuellar, S. Dronov, E. D. Gilles, M. Ginkel, V. Gor, I. I. Goryanin, W. J. Hedley, T. C. Hodgman, J.-H. Hofmeyr, P. J. Hunter, N. S. Juty, J. L. Kasberger, A. Kremling, U. Kummer, N. Le Novère, L. M. Loew, D. Lucio, P. Mendes, E. Minch, E. D. Mjolsness, Y. Nakayama, M. R. Nelson, P. F. Nielsen, T. Sakurada, J. C. Schaff, B. E. Shapiro, T. S. Shimizu, H. D. Spence, J. Stelling, K. Takahashi, M. Tomita, J. Wagner, J. Wang, S. Forum. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* 4 Mar. 2003, 19, 524–531.



- [35] P. Romero, J. Wagg, M. L. Green, D. Kaiser, M. Krummenacker, P. D. Karp. Computational prediction of human metabolic pathways from the complete human genome. *Genome biology* 1 **2005**, 6, R2.
- [36] T. Slater. Recent advances in modeling languages for pathway maps and computable biological networks. *Drug discovery today* 2 **Feb. 2014**, 19, 193–198.
- [37] K. Degtyarenko, P. de Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcántara, M. Darsow, M. Guedj, M. Ashburner. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic acids research Database issue* **Jan. 2008**, 36, D344–D350.
- [38] B. Braschi, P. Denny, K. Gray, T. Jones, R. Seal, S. Tweedie, B. Yates, E. Bruford. Genenames.org: the HGNC and VGNC resources in 2019. *Nucleic acids research D1* **Jan. 2019**, 47, D786–D792.
- [39] O. Consortium. *Biological Expression Language tools and services*. URL: <https://github.com/belbio/> (visited on 08/08/2019).
- [40] C. T. Hoyt, A. Konotopez, C. Ebeling, J. Wren. PyBEL: a computational framework for Biological Expression Language. *Bioinformatics (Oxford, England)* 4 **Feb. 2018**, 34, 703–704.
- [41] C. T. Hoyt, D. Domingo-Fernández, S. Mubeen, J. M. Llaó, A. Konotopez, C. Ebeling, C. Birkenbihl, Ö. Muslu, B. English, S. Müller, M. P. de Lacerda, M. Ali, S. Colby, D. Türei, N. Palacio-Escat, M. Hofmann-Apitius. Integration of Structured Biological Data Sources using Biological Expression Language. *bioRxiv* **2019**. eprint: <https://www.biorxiv.org/content/early/2019/05/08/631812.full.pdf>.
- [42] N. Le Novère, M. Hucka, H. Mi, S. Moodie, F. Schreiber, A. Sorokin, E. Demir, K. Wegner, M. I. Aladjem, S. M. Wimalaratne, F. T. Bergman, R. Gauges, P. Ghazal, H. Kawaji, L. Li, Y. Matsuoka, A. Villéger, S. E. Boyd, L. Calzone, M. Courtot, U. Dogrusoz, T. C. Freeman, A. Funahashi, S. Ghosh, A. Jouraku, S. Kim, F. Kolpakov, A. Luna, S. Sahle, E. Schmidt, S. Watterson, G. Wu, I. Goryanin, D. B. Kell, C. Sander, H. Sauro, J. L. Snoep, K. Kohn, H. Kitano. The Systems Biology Graphical Notation. *Nature biotechnology* 8 **Aug. 2009**, 27, 735–741.
- [43] S. Orchard, P. Kersey, H. Hermjakob, R. Apweiler. The HUPO Proteomics Standards Initiative Meeting: Towards Common Standards for Exchanging Proteomics Data. *Comparative and functional genomics* 1 **2003**, 4, 16–19.

- [44] A. K. Miller, J. Marsh, A. Reeve, A. Garny, R. Britten, M. Halstead, J. Cooper, D. P. Nickerson, P. F. Nielsen. An overview of the CellML API and its implementation. *BMC bioinformatics* **Apr. 2010**, 11, 178.
- [45] Y. Matsuoka, S. Ghosh, N. Kikuchi, H. Kitano. Payao: a community platform for SBML pathway model curation. *Bioinformatics* 10 **May 2010**, 26, 1381–1383.
- [46] P. Gawron, M. Ostaszewski, V. Satagopam, S. Gebel, A. Mazein, M. Kuzma, S. Zorzan, F. McGee, B. Otjacques, R. Balling, R. Schneider. MINERVA—a platform for visualization and curation of molecular interaction networks. *NPJ systems biology and applications* **2016**, 2, 16020.
- [47] T. Czauderna, C. Klukas, F. Schreiber. Editing, validating and translating of SBGN maps. *Bioinformatics* 18 **Sept. 2010**, 26, 2340–2341.
- [48] I. Kuperstein, D. P. A. Cohen, S. Pook, E. Viara, L. Calzone, E. Barillot, A. Zinovyev. NaviCell: a web-based environment for navigation, curation and maintenance of large molecular interaction maps. *BMC systems biology* **Oct. 2013**, 7, 100.
- [49] M. Kutmon, M. P. van Iersel, A. Bohler, T. Kelder, N. Nunes, A. R. Pico, C. T. Evelo. PathVisio 3: an extendable pathway analysis toolbox. *PLoS computational biology* 2 **Feb. 2015**, 11, e1004085.
- [50] E. Bonnet, E. Viara, I. Kuperstein, L. Calzone, D. P. A. Cohen, E. Barillot, A. Zinovyev. NaviCell Web Service for network-based data visualization. *Nucleic acids research* W1 **July 2015**, 43, W560–W565.
- [51] E. Bonnet, L. Calzone, D. Rovera, G. Stoll, E. Barillot, A. Zinovyev. BiNoM 2.0, a Cytoscape plugin for accessing and analyzing pathways using standard systems biology formats. *BMC systems biology* **Mar. 2013**, 7, 18.
- [52] R. T. Pillich, J. Chen, V. Rynkov, D. Welker, D. Pratt. NDEx: A Community Resource for Sharing and Publishing of Biological Networks. *Methods in molecular biology (Clifton, N.J.)* **2017**, 1558, 271–301.
- [53] D. Pratt, J. Chen, D. Welker, R. Rivas, R. Pillich, V. Rynkov, K. Ono, C. Miello, L. Hicks, S. Szalma, A. Stojmirovic, R. Dobrin, M. Braxenthaler, J. Kuentzer, B. Demchak, T. Ideker. NDEx, the Network Data Exchange. *Cell systems* 4 **Oct. 2015**, 1, 302–305.
- [54] D. Pratt, J. Chen, R. Pillich, V. Rynkov, A. Gary, B. Demchak, T. Ideker. NDEx 2.0: A Clearinghouse for Research on Cancer Pathways. *Cancer research* 21 **Nov. 2017**, 77, e58–e61.

- [55] Z. A. King, A. Dräger, A. Ebrahim, N. Sonnenschein, N. E. Lewis, B. O. Palsson. Escher: A Web Application for Building, Sharing, and Embedding Data-Rich Visualizations of Biological Pathways. *PLoS computational biology* 8 **Aug. 2015**, 11, e1004321.
- [56] C. Vehlow, J. Hasenauer, A. Kramer, A. Raue, S. Hug, J. Timmer, N. Radde, F. J. Theis, D. Weiskopf. iVUN: interactive Visualization of Uncertain biochemical reaction Networks. *BMC bioinformatics* **2013**, 14 Suppl 19, S2.
- [57] E. Demir, O. Babur, I. Rodchenkov, B. A. Aksoy, K. I. Fukuda, B. Gross, O. S. Sümer, G. D. Bader, C. Sander. Using biological pathway data with paxtools. *PLoS computational biology* 9 **2013**, 9, e1003194.
- [58] C. Wrzodek, A. Dräger, A. Zell. KEGGtranslator: visualizing and converting the KEGG PATHWAY database to various formats. *Bioinformatics* 16 **Aug. 2011**, 27, 2314–2315.
- [59] F. Kramer, M. Bayerlová, F. Klemm, A. Bleckmann, T. Beissbarth. rBiopaxParser—an R package to parse, modify and visualize BioPAX data. *Bioinformatics (Oxford, England)* 4 **Feb. 2013**, 29, 520–522.
- [60] B. M. Gyori, J. A. Bachman, K. Subramanian, J. L. Muhlich, L. Galescu, P. K. Sorger. From word models to executable models of signaling networks using automated assembly. *Molecular systems biology* 11 **Nov. 2017**, 13, 954.
- [61] M. D. Stobbe, S. M. Houten, G. A. Jansen, A. H. C. van Kampen, P. D. Moerland. Critical assessment of human metabolic pathway databases: a stepping stone for future integration. *BMC systems biology* **Oct. 2011**, 5, 165.
- [62] D. Türei, T. Korcsmáros, J. Saez-Rodriguez. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nature methods* 12 **Nov. 2016**, 13, 966–967.
- [63] G. Sales, E. Calura, C. Romualdi. metaGraphite - a new layer of pathway annotation to get metabolite networks. *Bioinformatics* **Sept. 2018**.
- [64] M. Nickel, L. Rosasco, T. Poggio. Holographic Embeddings of Knowledge Graphs. In: *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. AAAI'16. Phoenix, Arizona: AAAI Press, **2016**, 1955–1961.
- [65] M. Nickel, K. Murphy, V. Tresp, E. Gabrilovich. A Review of Relational Machine Learning for Knowledge Graphs. *Proceedings of the IJCAI* **2016**, (1), 11–33.

- [66] J. Zhou, G. Cui, Z. Zhang, C. Yang, Z. Liu, M. Sun. Graph Neural Networks: A Review of Methods and Applications. *CoRR* **2018**, abs/1812.08434. arXiv: 1812.08434.
- [67] M. Ostaszewski, S. Gebel, I. Kuperstein, A. Mazein, A. Zinovyev, U. Dogrusoz, J. Hasenauer, R. M. T. Fleming, N. Le Novère, P. Gawron, T. Ligon, A. Niarakis, D. Nickerson, D. Weindl, R. Balling, E. Barillot, C. Auffray, R. Schneider. Community-driven roadmap for integrated disease maps. *Briefings in bioinformatics* **Apr. 2018**.
- [68] P. Khatri, M. Sirota, A. J. Butte. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS computational biology* **2012**, *8*, e1002375.
- [69] M. Saqi, A. Lysenko, Y.-K. Guo, T. Tsunoda, C. Auffray. Navigating the disease landscape: knowledge representations for contextualizing molecular signatures. *Briefings in bioinformatics* **Apr. 2018**.
- [70] S. Ogishima, S. Mizuno, M. Kikuchi, A. Miyashita, R. Kuwano, H. Tanaka, J. Nakaya. AlzPathway, an Updated Map of Curated Signaling Pathways: Towards Deciphering Alzheimer's Disease Pathogenesis. *Methods in molecular biology* **2016**, *1303*, 423–432.
- [71] S. Mizuno, R. Iijima, S. Ogishima, M. Kikuchi, Y. Matsuoka, S. Ghosh, T. Miyamoto, A. Miyashita, R. Kuwano, H. Tanaka. AlzPathway: a comprehensive map of signaling pathways of Alzheimer's disease. *eng. BMC Syst Biol* **2012**, *6*, 52.
- [72] K. A. Fujita, M. Ostaszewski, Y. Matsuoka, S. Ghosh, E. Glaab, C. Trefois, I. Crespo, T. M. Perumal, W. Jurkowski, P. M. A. Antony, N. Diederich, M. Buttini, A. Kodama, V. P. Satagopam, S. Eifes, A. Del Sol, R. Schneider, H. Kitano, R. Balling. Integrating pathways of Parkinson's disease in a molecular interaction map. *eng. Mol Neurobiol* **Feb. 2014**, *49* (1), 88–102.
- [73] A. Mazein, R. G. Knowles, I. Adcock, K. F. Chung, C. E. Wheelock, A. H. Maitland-van der Zee, P. J. Sterk, C. Auffray, A. P. Team. AsthmaMap: An expert-driven computational representation of disease mechanisms. *Clinical and experimental allergy: journal of the British Society for Allergy and Clinical Immunology* **8 Aug. 2018**, *48*, 916–918.
- [74] I. Kuperstein, E. Bonnet, H.-A. Nguyen, D. Cohen, E. Viara, L. Grieco, S. Fourquet, L. Calzone, C. Russo, M. Kondratova, M. Dutreix, E. Barillot, A. Zinovyev. Atlas of Cancer Signalling Network: a systems biology resource for integrative analysis of cancer data with Google Maps. *Oncogenesis* **July 2015**, *4*, e160.

- [75] V. Singh, M. Ostaszewski, G. D. Kallioliias, G. Chiocchia, R. Olaso, E. Petit-Teixeira, T. Helikar, A. Niarakis. Computational Systems Biology Approach for the Study of Rheumatoid Arthritis: From a Molecular Map to a Dynamical Model. *Genomics and computational biology* 1 **2018**, 4.
- [76] Y. Matsuoka, H. Matsumae, M. Katoh, A. J. Einfeld, G. Neumann, T. Hase, S. Ghosh, J. E. Shoemaker, T. J. S. Lopes, T. Watanabe, S. Watanabe, S. Fukuyama, H. Kitano, Y. Kawaoka. A comprehensive map of the influenza A virus replication cycle. *BMC systems biology* **Oct. 2013**, 7, 97.
- [77] A. Mazein, M. Ostaszewski, I. Kuperstein, S. Watterson, N. Le Novère, D. Lefaudeux, B. De Meulder, J. Pellet, I. Balaur, M. Saqi, M. M. Nogueira, F. He, A. Parton, N. Lemonnier, P. Gawron, S. Gebel, P. Hainaut, M. Ollert, U. Dogrusoz, E. Barillot, A. Zinovyev, R. Schneider, R. Balling, C. Auffray. Systems medicine disease maps: community-driven comprehensive representation of disease mechanisms. *NPJ systems biology and applications* **2018**, 4, 21.
- [78] A. Wimo, M. Guerchet, G.-C. Ali, Y.-T. Wu, A. M. Prina, B. Winblad, L. Jönsson, Z. Liu, M. Prince. The worldwide costs of dementia 2015 and comparisons with 2010. *Alzheimer's & dementia: the journal of the Alzheimer's Association* 1 **Jan. 2017**, 13, 1–7.
- [79] C. M. Doran, I. Kinchin. A review of the economic impact of mental illness. *Australian health review: a publication of the Australian Hospital Association* 1 **Feb. 2019**, 43, 43–48.
- [80] T. Pringsheim, K. Fiest, N. Jette. The international incidence and prevalence of neurologic conditions: how common are they? *Neurology* 18 **Oct. 2014**, 83, 1661–1664.
- [81] D. Hirtz, D. J. Thurman, K. Gwinn-Hardy, M. Mohamed, A. R. Chaudhuri, R. Zalutsky. How common are the "common" neurologic disorders? *Neurology* 5 **Jan. 2007**, 68, 326–337.
- [82] I. Grundke-Iqbal, K. Iqbal, Y. C. Tung, M. Quinlan, H. M. Wisniewski, L. I. Binder. Abnormal phosphorylation of the microtubule-associated protein tau (tau) in Alzheimer cytoskeletal pathology. *Proceedings of the National Academy of Sciences of the United States of America* 13 **July 1986**, 83, 4913–4917.

- [83] M. T. Heneka, M. J. Carson, J. El Khoury, G. E. Landreth, F. Brosseron, D. L. Feinstein, A. H. Jacobs, T. Wyss-Coray, J. Vitorica, R. M. Ransohoff, K. Herrup, S. A. Frautschy, B. Finsen, G. C. Brown, A. Verkhratsky, K. Yamanaka, J. Koistinaho, E. Latz, A. Halle, G. C. Petzold, T. Town, D. Morgan, M. L. Shinohara, V. H. Perry, C. Holmes, N. G. Bazan, D. J. Brooks, S. Hunot, B. Joseph, N. Deigendesch, O. Garaschuk, E. Boddeke, C. A. Dinarello, J. C. Breitner, G. M. Cole, D. T. Golenbock, M. P. Kummer. Neuroinflammation in Alzheimer's disease. *The Lancet. Neurology* 4 **Apr. 2015**, 14, 388–405.
- [84] P. T. Francis, A. M. Palmer, M. Snape, G. K. Wilcock. The cholinergic hypothesis of Alzheimer's disease: a review of progress. *Journal of neurology, neurosurgery, and psychiatry* 2 **Feb. 1999**, 66, 137–147.
- [85] X. Du, X. Wang, M. Geng. Alzheimer's disease hypothesis and related therapies. *Translational neurodegeneration* **2018**, 7, 2.
- [86] M. P. Murphy, H. LeVine. Alzheimer's disease and the amyloid-beta peptide. *Journal of Alzheimer's disease: JAD* 1 **2010**, 19, 311–323.
- [87] A. T. Kodamullil, F. Zekri, M. Sood, B. Hengerer, L. Canard, D. McHale, M. Hofmann-Apitius. Trial watch: Tracing investment in drug development for Alzheimer disease. *Nature reviews. Drug discovery* 12 **Dec. 2017**, 16, 819.
- [88] D. Mehta, R. Jackson, G. Paul, J. Shi, M. Sabbagh. Why do trials for Alzheimer's disease drugs keep failing? A discontinued drug perspective for 2010-2015. *Expert opinion on investigational drugs* 6 **June 2017**, 26, 735–739.
- [89] A. D. International. *World Alzheimer Report 2018*. URL: <https://www.alz.co.uk/research/world-report-2018> (visited on 08/08/2019).
- [90] I. Piaceri, B. Nacmias, S. Sorbi. Genetics of familial and sporadic Alzheimer's disease. *Frontiers in bioscience* **Jan. 2013**, 5, 167–177.
- [91] J. Dorszewska, M. Prendecki, A. Oczkowska, M. Dezor, W. Kozubski. Molecular Basis of Familial and Sporadic Alzheimer's Disease. *Current Alzheimer research* 9 **2016**, 13, 952–963.
- [92] R. Duara, R. F. Lopez-Alberola, W. W. Barker, D. A. Loewenstein, M. Zatzinsky, C. E. Eisdorfer, G. B. Weinberg. A comparison of familial and sporadic Alzheimer's disease. *Neurology* 7 **July 1993**, 43, 1377–1384.
- [93] N. P. Foundation. *What Is Parkinson's*. URL: <http://www.parkinson.org/understanding-parkinsons/what-is-parkinsons> (visited on 05/02/2019).

- [94] J. Xu, D. D. Gong, C. F. Man, Y. Fan. Parkinson's disease and risk of mortality: meta-analysis and systematic review. eng. *Acta Neurol Scand* **Feb. 2014**, 129 (2), 71–79.
- [95] W. A. Rocca. The burden of Parkinson's disease: a worldwide perspective. *The Lancet. Neurology* 11 **Nov. 2018**, 17, 928–929.
- [96] Alzheimers.net. *Alzheimer's statistics*. URL: <http://www.alzheimers.net/resources/alzheimers-statistics/> (visited on 05/02/2019).
- [97] G. F. Wooten, L. J. Currie, V. E. Bovbjerg, J. K. Lee, J. Patrie. Are men at greater risk for Parkinson's disease than women? *Journal of neurology, neurosurgery, and psychiatry* 4 **Apr. 2004**, 75, 637–639.
- [98] C. A. Haaxma, B. R. Bloem, G. F. Borm, W. J. G. Oyen, K. L. Leenders, S. Eshuis, J. Booij, D. E. Dluzen, M. W. I. M. Horstink. Gender differences in Parkinson's disease. *Journal of neurology, neurosurgery, and psychiatry* 8 **Aug. 2007**, 78, 819–824.
- [99] A. I. Rodriguez-Perez, A. Borrajo, J. Rodriguez-Pallares, M. J. Guerra, J. L. Labandeira-Garcia. Interaction between NADPH-oxidase and Rho-kinase in angiotensin II-induced microglial activation. *Glia* 3 **Mar. 2015**, 63, 466–482.
- [100] M. P. Helley, J. Pinnell, C. Sportelli, K. Tieu. Mitochondria: A Common Target for Genetic Mutations and Environmental Toxicants in Parkinson's Disease. *Frontiers in genetics* **2017**, 8, 177.
- [101] D. T. Dexter, P. Jenner. Parkinson disease: from pathology to molecular disease mechanisms. eng. *Free Radic Biol Med* **Sept. 2013**, 62, 132–144.
- [102] D. Vibha, S. Sureshbabu, G. Shukla, V. Goyal, A. K. Srivastava, S. Singh, M. Behari. Differences between familial and sporadic Parkinson's disease. *Parkinsonism & related disorders* 7 **Aug. 2010**, 16, 486–487.
- [103] S. Lesage, A. Brice. Parkinson's disease: from monogenic forms to genetic susceptibility factors. eng. *Hum Mol Genet* **Apr. 2009**, 18 (R1), R48–R59.
- [104] T. Kitada, J. J. Tomlinson, H. S. Ao, D. A. Grimes, M. G. Schlossmacher. Considerations regarding the etiology and future treatment of autosomal recessive versus idiopathic Parkinson disease. eng. *Curr Treat Options Neurol* **June 2012**, 14 (3), 230–240.

- [105] R. H. Pietrzak, R. B. Goldstein, S. M. Southwick, B. F. Grant. Psychiatric comorbidity of full and partial posttraumatic stress disorder among older adults in the United States: results from wave 2 of the National Epidemiologic Survey on Alcohol and Related Conditions. *The American journal of geriatric psychiatry: official journal of the American Association for Geriatric Psychiatry* 5 **May 2012**, 20, 380–390.
- [106] A. P. Association. Diagnostic and statistical manual of mental disorders: DSM-5. 5th ed. American Psychiatric Association, **2013**.
- [107] C. R. Bailey, E. Cordell, S. M. Sobin, A. Neumeister. Recent progress in understanding the pathophysiology of post-traumatic stress disorder: implications for targeted pharmacological treatment. *CNS drugs* 3 **Mar. 2013**, 27, 221–232.
- [108] R. L. Spitzer, K. K. Md, J. B. W. Williams. Diagnostic and Statistical Manual of Mental Disorders, Third Edition. 3th ed. American Psychiatric Association, **1980**.
- [109] R. C. Kessler, S. Aguilar-Gaxiola, J. Alonso, C. Benjet, E. J. Bromet, G. Cardoso, L. Degenhardt, G. de Girolamo, R. V. Dinolova, F. Ferry, S. Florescu, O. Gureje, J. M. Haro, Y. Huang, E. G. Karam, N. Kawakami, S. Lee, J.-P. Lepine, D. Levinson, F. Navarro-Mateu, B.-E. Pennell, M. Piazza, J. Posada-Villa, K. M. Scott, D. J. Stein, M. Ten Have, Y. Torres, M. C. Viana, M. V. Petukhova, N. A. Sampson, A. M. Zaslavsky, K. C. Koenen. Trauma and PTSD in the WHO World Mental Health Surveys. *European journal of psychotraumatology* sup5 **2017**, 8, 1353383.
- [110] C. Benjet, E. Bromet, E. G. Karam, R. C. Kessler, K. A. McLaughlin, A. M. Ruscio, V. Shahly, D. J. Stein, M. Petukhova, E. Hill, J. Alonso, L. Atwoli, B. Bunting, R. Bruffaerts, J. M. Caldas-de-Almeida, G. de Girolamo, S. Florescu, O. Gureje, Y. Huang, J. P. Lepine, N. Kawakami, V. Kovess-Masfety, M. E. Medina-Mora, F. Navarro-Mateu, M. Piazza, J. Posada-Villa, K. M. Scott, A. Shalev, T. Slade, M. ten Have, Y. Torres, M. C. Viana, Z. Zarkov, K. C. Koenen. The epidemiology of traumatic event exposure worldwide: results from the World Mental Health Survey Consortium. *Psychological medicine* 2 **Jan. 2016**, 46, 327–343.
- [111] F. R. Ferry, S. E. Brady, B. P. Bunting, S. D. Murphy, D. Bolton, S. M. O'Neill. The Economic Burden of PTSD in Northern Ireland. *Journal of traumatic stress* 3 **June 2015**, 28, 191–197.
- [112] J. M. Ferguson. SSRI Antidepressant Medications: Adverse Effects and Tolerability. *Journal of clinical psychiatry* 1 **Feb. 2001**, 3, 22–27.



- [113] X. Zhang, X.-F. Sun, Y. Cao, B. Ye, Q. Peng, X. Liu, B. Shen, H. Zhang. CBD: a biomarker database for colorectal cancer. *Database: the journal of biological databases and curation* **Jan. 2018**, 2018.
- [114] T. Clark, J. Kinoshita. Alzforum and SWAN: the present and future of scientific web communities. *Briefings in bioinformatics* 3 **May 2007**, 8, 163–171.
- [115] S. Yerlikaya, T. Broger, E. MacLean, M. Pai, C. M. Denking. A tuberculosis biomarker database: the key to novel TB diagnostics. *International journal of infectious diseases: IJID: official publication of the International Society for Infectious Diseases* **Mar. 2017**, 56, 253–257.
- [116] H.-J. Dai, J. C.-Y. Wu, W.-S. Lin, A. J. F. Reyes, M. A. C. Dela Rosa, S. Syed-Abdul, R. T.-H. Tsai, W.-L. Hsu. LiverCancerMarkerRIF: a liver cancer biomarker interactive curation system combining text mining and expert annotations. *Database: the journal of biological databases and curation* **2014**, 2014.
- [117] M. A. García-Campos, J. Espinal-Enríquez, E. Hernández-Lemus. Pathway Analysis: State of the Art. *Frontiers in physiology* **2015**, 6, 383.
- [118] A. Fabregat, K. Sidiropoulos, P. Garapati, M. Gillespie, K. Hausmann, R. Haw, B. Jassal, S. Jupe, F. Korninger, S. McKay, L. Matthews, B. May, M. Milacic, K. Rothfels, V. Shamovsky, M. Webber, J. Weiser, M. Williams, G. Wu, L. Stein, H. Hermjakob, P. D'Eustachio. The Reactome pathway Knowledgebase. *Nucleic acids research D1* **Jan. 2016**, 44, D481–D487.
- [119] D. Domingo-Fernández, C. T. H. Hoyt. *ComPath Python Package*. **Oct. 2018**.
- [120] D. Domingo-Fernández, C. T. Hoyt, J. Marin, S. Mubeen. *PathMe*. **Feb. 2019**.
- [121] D. Domingo-Fernández, C. T. Hoyt, S. Mubeen. *PathMe Viewer*. **2019**.
- [122] L. Wadi, M. Meyer, J. Weiser, L. D. Stein, J. Reimand. Impact of outdated gene annotations on pathway enrichment analysis. *Nature methods* 9 **Aug. 2016**, 13, 705–706.
- [123] C. T. Hoyt, D. Domingo-Fernández, N. Balzer, A. Güldenpfennig, M. Hofmann-Apitius. A systematic approach for identifying shared mechanisms in epilepsy and its comorbidities. *Database: the journal of biological databases and curation* **Jan. 2018**, 2018.
- [124] P. W. Tuerk, S. A. M. Rauch, B. O. Rothbaum. Effect size matters: a key neglected indicator of comparative trial quality. *The lancet. Psychiatry* 2 **Feb. 2019**, 6, e4.

- [125] I. C. Passos, M. P. Vasconcelos-Moreno, L. G. Costa, M. Kunz, E. Brietzke, J. Quevedo, G. Salum, P. V. Magalhães, F. Kapczinski, M. Kauer-Sant'Anna. Inflammatory markers in post-traumatic stress disorder: a systematic review, meta-analysis, and meta-regression. *The lancet. Psychiatry* 11 **Nov. 2015**, 2, 1002–1012.
- [126] S. Nidich, P. J. Mills, M. Rainforth, P. Heppner, R. H. Schneider, N. E. Rosenthal, J. Salerno, C. Gaylord-King, T. Rutledge. Non-trauma-focused meditation versus exposure therapy in veterans with post-traumatic stress disorder: a randomised controlled trial. *The lancet. Psychiatry* 12 **Dec. 2018**, 5, 975–986.
- [127] M. Grassi, G. Perna, D. Caldirola, K. Schruers, R. Duara, D. A. Loewenstein. A Clinically-Translatable Machine Learning Algorithm for the Prediction of Alzheimer's Disease Conversion in Individuals with Mild and Premild Cognitive Impairment. *Journal of Alzheimer's disease: JAD* 4 **2018**, 61, 1555–1573.
- [128] G. Lee, K. Nho, B. Kang, K.-A. Sohn, D. Kim, for Alzheimer's Disease Neuroimaging Initiative. Predicting Alzheimer's disease progression using multi-modal deep learning approach. *Scientific reports* 1 **Feb. 2019**, 9, 1952.
- [129] Y. Ding, J. H. Sohn, M. G. Kawczynski, H. Trivedi, R. Harnish, N. W. Jenkins, D. Lituiev, T. P. Copeland, M. S. Aboian, C. Mari Aparici, S. C. Behr, R. R. Flavell, S.-Y. Huang, K. A. Zalocusky, L. Nardo, Y. Seo, R. A. Hawkins, M. Hernandez Pampaloni, D. Hadley, B. L. Franc. A Deep Learning Model to Predict a Diagnosis of Alzheimer Disease by Using F-FDG PET of the Brain. *Radiology* 2 **Feb. 2019**, 290, 456–464.
- [130] P. O. Larsen, M. von Ins. The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics* 3 **Sept. 2010**, 84, 575–603.
- [131] S. Mangul, T. Mosqueiro, R. J. Abdill, D. Duong, K. Mitchell, V. Sarwal, B. Hill, J. Brito, R. J. Littman, B. Statz, A. K.-M. Lam, G. Dayama, L. Grieneisen, L. S. Martin, J. Flint, E. Eskin, R. Blekhman. Challenges and recommendations to improve the installability and archival stability of omics computational tools. *PLoS biology* 6 **June 2019**, 17, e3000333.
- [132] B. A. Grüning, S. Lampa, M. Vaudel, D. Blankenberg. Software engineering for scientific big data analysis. *GigaScience* 5 **May 2019**, 8.
- [133] M. Ali, C. T. Hoyt, D. Domingo-Fernández, J. Lehmann, H. Jabeen. BioKEEN: A library for learning and evaluating biological knowledge graph embeddings. *Bioinformatics* **Feb. 2019**.

- [134] A. T. Kodamullil, E. Younesi, M. Naz, S. Bagewadi, M. Hofmann-Apitius. Computable cause-and-effect models of healthy and Alzheimer's disease states and their mechanistic differential analysis. *Alzheimer's & dementia: the journal of the Alzheimer's Association* 11 **Nov. 2015**, 11, 1329–1339.
- [135] S. Picart-Armada, F. Fernández-Albert, M. Vinaixa, M. A. Rodríguez, S. Aivio, T. H. Stracker, O. Yanes, A. Perera-Lluna. Null diffusion-based enrichment for metabolomics data. *PloS one* 12 **2017**, 12, e0189012.
- [136] C. T. Hoyt, D. Domingo-Fernández, M. Hofmann-Apitius. BEL Commons: an environment for exploration and analysis of networks encoded in Biological Expression Language. *Database: the journal of biological databases and curation* **Jan. 2018**, 2018.