Generation and Applications of Knowledge Graphs in Systems and Networks Biology

Kumulative Dissertation

zur Erlangung des Doktorgrades (Dr. rer. nat.)

der Mathematisch-Naturwissenschaftlichen Fakultät

der Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

CHARLES TAPLEY HOYT

aus New Haven, United States of America

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: XX. September 2019
Erscheinungsjahr: 2019

Abstract

This thesis begins with the acknowledgement of the acceleration of generation of knowledge within the biomedical domain. The first two papers (PyBEL and BEL Commons) build an ecosystem for handling this knowledge during curation, application of algorithmics, and visualization. The second two papers revolve around enabling the acquisition of high-granularity knowledge from structured sources on a massive scale (Bio2BEL) and supporting the semi-automated curation of new content at high speed and precision (Re-curation and Rational Enrichment). Finally, after building the ecosystem and acquiring the content, the third part of this thesis revolves around the applications of biological knowledge graphs in simulation and modeling. This includes agent-based modeling using biological knowledge graphs as priors (BEL2ABM), the application of network representation learning to prioritize nodes in biological knowledge graphs based on corresponding experimental measurements (GuiltyTargets), and finally, the use of biological knowledge graphs and development of algorithmics to deconvolute the mechanism of action of drugs, that could also serve as a drug repositioning candidate identifier (EpiCom). Ultimately, the this thesis lays the groundwork for production-level applications of drug repositioning algorithms and other knowledge-driven approaches to analyzing biomedical experiments.

Acknowledgment

Martin

Supportive Department

Family

Daniel

Scott

Students

Publications

Thesis Publications

1. **Hoyt, C. T.**, Konotopez, A., & Ebeling, C. (2018). PyBEL: a computational framework for Biological Expression Language. *Bioinformatics (Oxford, England)*, 34(4), 703–704.

2. **Hoyt, C. T.**, Domingo-Fernández, D., & Hofmann-Apitius, M. (2018). BEL Commons: an environment for exploration and analysis of networks encoded in Biological Expression Language. *Database : The Journal of Biological Databases and Curation*, 2018(3), 1–11.

3. **Hoyt, C. T.**, et al. (2019). Re-curation and Rational Enrichment of Knowledge Graphs in Biological Expression Language. *bioRxiv*, 536409.

4. **Hoyt, C. T.**, et al. (2019). Bio2BEL: Integration of Structured Knowledge Sources with Biological Expression Language. *bioRxiv*, 536409.

5. Gündel, M., **Hoyt, C. T.**, & Hofmann-Apitius, M. (2018). BEL2ABM: Agent-based simulation of static models in Biological Expression Language. *Bioinformatics*, 34(13), 2316–2318.

6. **Hoyt, C. T.**, et al. (2018). A systematic approach for identifying shared mechanisms in epilepsy and its comorbidities. *Database : The Journal of Biological Databases and Curation*, 2018(1).

7. Muslu, Ö., **Hoyt, C. T.**, & Hofmann-Apitius, M., & Fröhlich, H. (2019). GuiltyTargets: Prioritization of Novel Therapeutic Targets with Deep Network Representation Learning. *bioRxiv*, 1–14.

Other Publications

- Bradford, R., Sturm, T., Weber, A., Davenport, J. H., England, M., Errami, H., Gerdt, V., Grigoriev, D., **Hoyt, C. T.**, Košta, M., & Radulescu, O. (2017). A Case Study on the Parametric Occurrence of Multiple Steady States. In Proceedings of the 2017 ACM on International Symposium on Symbolic and Algebraic Computation - ISSAC '17 (Vol. Part F1293, pp. 45–52). New York, New York, USA: ACM Press.
- Domingo-Fernández, D., **Hoyt, C. T.**, Alvarez, C. B., Marin-Llao, J., & Hofmann-Apitius, M. (2018). ComPath: an ecosystem for exploring, analyzing, and curating mappings across pathway databases. *Npj Systems Biology and Applications*, 5(1), 3.
- Domingo-Fernández, D., Mubeen, S., Marín-Llaó, J., **Hoyt, C. T.**, & Hofmann-Apitius, M. (2019). PathMe: merging and exploring mechanistic pathway knowledge. *BMC Bioinformatics*, 20(1), 243.
- Ali, M., **Hoyt, C. T.**, Domingo-Fernández, D., Lehmann, J., & Jabeen, H. (2019). BioKEEN: A library for learning and evaluating biological knowledge graph embeddings. *Bioinformatics (Oxford, England)*.
- Bradford, R., Davenport, J. H., England, M., Errami, H., Gerdt, V., Grigoriev, D., **Hoyt, C. T.**, Kosta, M., Radulescu, O., Sturm, T., & Weber, A. (2019). Identifying the Parametric Occurrence of Multiple Steady States for some Biological Networks. Retrieved from http://arxiv.org/abs/1902.04882

Contents

1 Introduction

1.1 Nomenclature

1.1.1 Issues with Gene Nomenclature

The nomenclature of genes and gene products is a particularly egregious example of nomenclature within the biomedical domain. Genes often have several names as well as several incomprehensible acronyms, or gene symbols. For example, HGNC [**Yates2017**] and Entrez Gene [**Maglott2011**] list that the human gene, microtubule associated protein tau (hgnc:HGNC:6893, ncbigene:4137), has previously been named the G protein $\beta 1/\gamma 2$ subunit-interacting factor 1 and the protein phosphatase 1, regulatory subunit 103. Like with many genes, it is often acronymized to MAPT in text, but it has additionally been previously referenced with DDPAC, FLJ31424, FTDP-17, MAPTL, MGC13854, MTBT1, MTBT2, MSTD, PPND, and PPP1R103.

Neither genes' names nor their gene symbols convey their host species, which leads to further ambiguities in articles discussing orthologs in model organisms. The organization responsible for mouse gene nomenclature, Mouse Genome Informatics (MGI) [**Blake2017**], names the mouse orthologous gene as microtubule-associated protein tau (mgi:MGI:97180, ncbigene:17762) and lists the gene symbol as Mapt. In this example, the name varies from the human ortholog with the introduction of a dash between "microtubule" and "associated." The gene symbol differs only in capitalization. Similarly, the organization for rat genome nomenclature, the Rat Genome Database (RGD) [**Shimoyama2015**], names the rat orthol-

Organism	Database	Reference
cell4	cell5	cell6
cell7	cell8	cell9

Table 1: A non-exhaustive list of model organism gene nomenclature databases

ogous gene as microtubule-associated protein tau (rgd:69329; ncbigene:29477) - exactly as in MGI. While these orthologs from common model organisms have had related names, organisms with genetic drift such as Zebrafish have several orthologs named microtubule-associated protein tau a (zfin:ZDB-GENE-081027-1) and microtubule-associated protein tau b (zfin:ZDB-GENE-081027-2) whose gene symbols are listed as mapta and maptb, respectively. Other orthologs to human microtubule-associated protein tau can be found in Homologene (homologene:74962), Ensembl [**Zerbino2018**], HGNC, MGI, PomBase, RGD, Xenbase, and ZFIN. The HGNC Comparison of Orthology Predictions (HCOP) [**Wright2005**] aggregates these and several other sources of curated and predicted orthologies.

1.1.2 Issues with Gene Nomenclature

Most biologically relevant named entities go by many names. For example, many genes were discovered and characterized in different labs and therefore named differently. As resources for exchanging genomic and protein sequences have become more ubiquitous in the last thirty years, it has become easier to reduce those duplicates. However, this does not solve the problem of establishing a canonical name for each entity. As alluded to in the previous section, several committees and consortia have formed to standardize the nomenclature used for genes for each species (Table 1).

2 PyBEL: a computational framework for Biological Expression Language

Introduction

Bioinformatics, 34(4), 2018, 703–704
doi: 10.1093/bioinformatics/btx660
Advance Access Publication Date: 18 October 2017
Applications Note

OXFORD

Systems biology

PyBEL: a computational framework for Biological Expression Language

Charles Tapley Hoyt[1,2,*], Andrej Konotopez[1] and Christian Ebeling[1]

[1]Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Sankt Augustin 53754, Germany and [2]Department of Life Science Informatics, Bonn-Aachen International Center for IT, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn 53113, Germany

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on July 5, 2017; revised on September 28, 2017; editorial decision on October 16, 2017; accepted on October 17, 2017

Abstract

Summary: Biological Expression Language (BEL) assembles knowledge networks from biological relations across multiple modes and scales. Here, we present PyBEL; a software package for parsing, validating, converting, storing, querying, and visualizing networks encoded in BEL.
Availability and implementation: PyBEL is implemented in platform-independent, universal Python code. Its source is distributed under the Apache 2.0 License at https://github.com/pybel.
Contact: charles.hoyt@scai.fraunhofer.de
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Among the most popular modeling and data exchange languages in systems biology are currently the Biological Pathways Exchange (BioPAX), Systems Biology Markup Language (SBML) and Biological Expression Language (BEL). BioPAX captures metabolic, signaling, molecular, gene-regulatory, and genetic interaction networks (Hucka *et al.*, 2003); SBML accommodates mathematical models of biochemical networks, cellular signaling, and metabolic pathways (Demir *et al.*, 2010); and BEL assembles qualitative causal and correlative relations between biological entities across multiple modes and scales, with full provenance information including namespace references, relation provenance (citation and evidence), and biological context-specific relation metadata (anatomy, cell, disease etc.) (Slater, 2014).

Although there exist several software packages for BioPAX and SBML, the ecosystem of open-source software for BEL is much more limited. An assessment of previous software (see Supplementary Table S3) shows there is an unmet need for easily installable, stable, facile software that parses modern BEL and provides programmatic access to a data container that enables the resulting network to be extended, queried, manipulated, analyzed, and visualized. Furthermore, a converter between common data formats is needed to enable re-usability and interoperability between general and BEL-specific software for network analysis and visualization.

Here, we present PyBEL; a software package designed to fulfill each of these needs.

2 Software architecture

The PyBEL software package consists of five main components: (i) network data container, (ii) parser and validator, (iii) network database manager, (iv) data converter and (v) network visualizer.

Although a graph refers to an abstraction for a set of objects (i.e. nodes) and their relations (i.e. edges), its instantiation in a real-world application is often called a network. We provide an implementation of a directed multigraph (i.e. a graph whose edges have directionality and any given pair of nodes may have multiple edges) that maps the biological entities and concepts in the subjects and objects of BEL relations to nodes in a network and their relations, with corresponding metadata, to edges. We extended the MultiDiGraph class from `NetworkX` (http://networkx.github.io) to enable users direct access to their suite of network algorithms and static visualizations to support their further development into biologically meaningful analyses.

The parser performs tokenization, lexical analysis, parsing, and validation on each of the three sections of BEL documents (see Supplementary Figs S1 and S2). Callbacks are used to annotate the entries in the document metadata section to a network instance, download and store the resources referenced in the definitions section,

maintain a list of current annotations from SET statements, and parse BEL relations to populate a network instance with the corresponding nodes, edges, and their metadata from the current internal state. Although relations' syntax is implicitly validated, the semantics of their subjects' and objects' identifiers are validated against the references from the definitions section. Finally, feedback is provided to users to support thoughtful re-curation, which could lead to more robust knowledge assemblies and enable more reproducible science.

Namespaces and networks are cached with a relational database to improve the speed of validation and access to data. Although relational databases lack the faculty for applying network algorithms, they provide indexing functionality that enables complicated queries and filters over the nodes, edges, and metadata of increasingly large collections of networks. For example, this could help identify intersections and potential cross-talk between disease-specific networks.

We implemented lossless converters for common file formats including Node-Link JSON, JGIF, CX, and binary as well as for database formats including SQL, Neo4J, and NDEx. We also provide lossy exporters to Excel, CSV, SIF, XGMML, and GSEA to facilitate usage in other programs. Notably, we have deferred implementing a RDF (Resource Description Framework) converter until improvements are made to the existing BEL to RDF mapping and its documentation (https://wiki.openbel.org). Future work will also include converters for BioPAX and SBML. See Supplementary Tables S1 and S2 for more detailed descriptions of each format.

Networks can be exported for visualization in Cytoscape or uploaded to NDEx (Pratt *et al.*, 2015) to take advantage of its viewer and simple query interface. Alternatively, we provide an interactive network explorer tailored to BEL networks (appropriate node coloring, metadata pop-ups etc.) that can be directly embedded as HTML in email, Jupyter Notebook, or a web application. It has already been used to produce visualizations in the NeuroMMSig Web Service (Domingo-Fernández *et al.*, 2017). Supplementary Figures S3–S5 present these visualizations side-by-side. In addition to their programmatic interfaces, the parser, storage, conversion, and visualization features are exposed via a command line tool.

3 Case study

The PyBEL suite includes functions for querying and mutating networks with which it implements state-of-the-art algorithms for over-representation analysis, functional class scoring, and pathway topological analysis of BEL networks such as Reverse Causal Reasoning (Catlett *et al.*, 2013). Figure 1 presents a case study in which a novel heat diffusion work flow was used to assess the observed impact on biological processes from differential gene expression in Alzheimer's disease (AD). Technical documentation is included in the Supplementary Material.

4 Discussion

Even after its v2.0 update, BEL does not yet explicitly specify many concepts in molecular biology such as epigenetic information (Irin *et al.*, 2015). The inevitability of language evolution prompted us to develop the parser in modules so that new syntax could be proposed and implemented quickly. As a proof of concept, a syntax extension for gene modifications is included in the package by default.

Historically, BEL has used a custom namespace file format, but the creation and maintenance of biological terminologies has tended towards using OWL (Web Ontology Language). Furthermore, many domains (e.g. SNPs) are growing too large to enumerate during semantic integration and validation. The modular architecture of the parser enables easy implementation of new definition file formats, external validation services, or even alternative namespace definition schemes to address these issues.

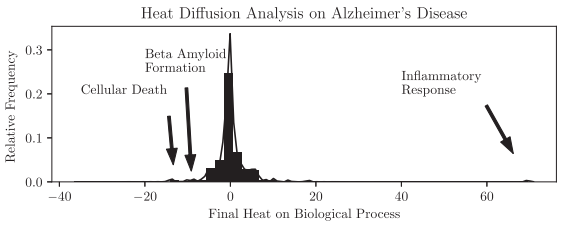


Fig. 1. Plotted is the distribution of the final heat on biological processes from the NeuroMMSig AD Knowledge Assembly (Domingo-Fernández *et al.*, 2017) following heat diffusion analysis with a differential gene expression experiment from the brains of AD patients (E-GEOD-5281, Liang *et al.*, 2007). The significant down-regulation of biological processes related to inflammatory response (heat = 69) and up-regulation of cellular death (heat = −13) and beta-amyloid formation (heat = −9) match common clinical observations and serve as a validation for this approach

Although BEL is often used to formalize knowledge curated from unstructured sources, our software also enables the integration of knowledge from structured sources. For example, existing solutions for resolving equivalences across namespaces rely on the creation and hosting of extensive lookup tables. Alternatively, the parser could be extended with a dedicated syntax and draw equivalencies directly from OWL.

Finally, we plan to present this software as a web service to enable a wider audience of researchers across disciplines to validate, explore, and analyze their BEL networks.

Acknowledgements

We thank Sumit Madan and Scott Colby for their advice and feedback.

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

Catlett,N. *et al.* (2013) Reverse causal reasoning: applying qualitative causal knowledge to the interpretation of high-throughput data. *BMC Bioinformatics*, **14**, 340.

Demir,E. *et al.* (2010) The BioPAX community standard for pathway data sharing. *Nat. Biotechnol.*, **28**, 1308–1308.

Domingo-Fernández,D. *et al.* (2017) Multimodal Mechanistic Signatures for Neurodegenerative Diseases (NeuroMMSig): a web server for mechanism enrichment. *Bioinformatics (Oxford, England)*, btx399.

Hucka,M. *et al.* (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics (Oxford, England)*, **19**, 524–531.

Irin,A.K. *et al.* (2015) Computational Modelling Approaches on Epigenetic Factors in Neurodegenerative and Autoimmune Diseases and Their Mechanistic Analysis. *J. Immunol. Res.*, **2015**, doi:10.1155/2015/737168.

Liang,W.S. *et al.* (2007) Gene expression profiles in anatomically and functionally distinct regions of the normal aged human brain. *Physiol. Genomics*, **28**, 311–322.

Pratt,D. *et al.* (2015) NDEx, the network data exchange. *Cell Syst.*, **1**, 302–305.

Slater,T. (2014) Recent advances in modeling languages for pathway maps and computable biological networks. *Drug Discov. Today*, **19**, 193–198.

Conclusions

3 BEL Commons: an environment for exploration and analysis of networks encoded in Biological Expression Language

Introduction

DATABASE
The Journal of Biological Databases and Curation

Database, 2018, 1–11
doi: 10.1093/database/bay126
Original article

Original article

BEL Commons: an environment for exploration and analysis of networks encoded in Biological Expression Language

Charles Tapley Hoyt[1,2,*], Daniel Domingo-Fernández[1,2] and Martin Hofmann-Apitius[1,2]

[1]Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing, 53754 Sankt Augustin, Germany and [2]Bonn-Aachen International Center for IT, Rheinische Friedrich-Wilhelms-Universität Bonn, 53115 Bonn, Germany

*Corresponding author: Tel: +49 2241 14 2268; Fax: +49 2241 14-2656; Email: charles.hoyt@scai.fraunhofer.de

Citation details: Hoyt,C. T., Domingo-Fernández,D. and Hofmann-Apitius,M. BEL Commons: an environment for exploration and analysis of networks encoded in Biological Expression Language. *Database* (2018) Vol. 2018: article ID bay126; doi:10.1093/database/bay126

Received 3 April 2018; Revised 25 July 2018; Accepted 5 November 2018

Abstract

The rapid accumulation of knowledge in the field of systems and networks biology during recent years requires complex, but user-friendly and accessible web applications that allow from visualization to complex algorithmic analysis. While several web applications exist with various focuses on creation, revision, curation, storage, integration, collaboration, exploration, visualization and analysis, many of these services remain disjoint and have yet to be packaged into a cohesive environment.

Here, we present BEL Commons: an integrative knowledge discovery environment for networks encoded in the Biological Expression Language (BEL). Users can upload files in BEL to be parsed, validated, compiled and stored with fine granular permissions. After, users can summarize, explore and optionally shared their networks with the scientific community. We have implemented a query builder wizard to help users find the relevant portions of increasingly large and complex networks and a visualization interface that allows them to explore their resulting networks. Finally, we have included a dedicated analytical service for performing data-driven analysis of knowledge networks to support hypothesis generation.

Database URL: https://bel-commons.scai.fraunhofer.de

Introduction

There exists a variety of modeling languages, data formats and analytical tools for systems and networks biology. Among the most popular modeling languages are the Biological Pathways Exchange [BioPAX, (1)], Systems Biology Markup Language [SBML, (2)], Systems Biology Graphical Notation [SBGN, (3)] and Biological Expression Language [BEL, (4)]. BioPAX captures metabolic, signaling,

molecular, gene-regulatory and genetic interaction networks; SBML captures mathematical models of biochemical networks, cellular signaling and metabolic pathways; SBGN provides a graphical representation of ideas from BioPAX and SBML; and BEL captures qualitative causal and correlative relations between biological entities across multiple scales (e.g. *-omics*, pathway, cellular, phenotypic) with accompanying biological and contextual annotations. While each modeling language has their own domain-specific syntax and semantics, each facilitates assembling biological relations into networks. For example, BEL formalizes relations as triplets each composed of a subject, predicate and object in order to generate pathways and networks when the object from one relation is again used as the subject of another. We refer to Saqi *et al.* (5) for a more thorough comparison of the applicabilities of various modeling languages.

Currently, these modeling languages and their related analytical tools require deep knowledge of computer programming to use and are generally inaccessible to a wider audience of biologists and clinicians. With the explosion of data and knowledge in the biomedical domain, it is paramount to develop tools that foster collaboration between groups of scientists with different backgrounds and skill sets who are working toward similar goals. Already, there are multiple freely available, web-based tools for systems and networks biology with varying focuses on creation, revision, curation, storage, integration, collaboration, exploration, visualization and analysis. Below, we provide a brief review of services appropriate for each focus.

Because of the accelerating throughput of scientific publication in the biomedical domain, several general workflows [e.g. Reading and Assembling Contextual and Holistic Mechanisms from Text (REACH) (6), TRIPS (7), Turku Event Extraction System (TEES) (8), MedScan (9)] and several BEL-specific workflows [e.g. Biological Expression Language Information Extraction WorkFlow (BELIEF) (10), BELMiner (11), BelSmile (12), BELTracker (13)] have been developed to automate biological relation extraction. The limits of the precision and recall of automated techniques, the applicability domains of different modeling languages and the need for expert input motivated the development of semi-automatic and manual curation interfaces [e.g. SBV Improver (14), BELIEF Dashboard (15), WikiPathways (16)] to drive crowdsourced creation, revision and curation of knowledge networks.

While many useful knowledge resources [e.g. miRTarBase (17), Comparative Toxicogenomics Database (CTD) (18), Kyoto Encyclopedia of Genes and Genomes (KEGG) (19)] still disseminate their data in non-standard formats, the use of the aforementioned modeling languages has become much more common as in the case of the integration effort of Pathway Commons (20). Other web tools [e.g. NDEx (21), GraphSpace (22)] provide the ability to upload, store, share and distribute networks while remaining agnostic to format. Most of these resources also include network visualization, layout and exploration with PathVisio (23), Cytoscape (24), Cytoscape.js (25), as well as in-browser navigators.

Numerous algorithms and analyses have been published for systems and networks biology but most are bespoke due to the heterogeneous nature of their underlying knowledge networks, data sets and the scientific questions motivating their development. An exception lies with gene set enrichment analysis: a technique for finding gene sets, pathways and networks in which a query gene set (e.g. a list of differentially expressed genes) is over- or under-represented (26). Its simplicity has led to its implementation and inclusion in several web applications as well as several applications to reveal patterns of dysregulation in *-omics* data sets as exemplified by the Enrichment Map Cytoscape Plugin (27, 28) with Pathway Commons as well as Gene Set Enrichment Analysis (GSEA) (29) with MSigDB (30).

Recently, BEL has been successfully used as a semantic and modeling framework for multi-scale and multimodal knowledge in order to investigate the etiology of complex neurodegenerative diseases as shown by Domingo-Fernández *et al.* (31) with the NeuroMMSig Mechanism Enrichment Server. While the list of published BEL-specific algorithms is currently short (e.g. Reverse Causal Reasoning (32), Network Perturbation Amplitude (33) etc.), recent developments in the BEL software ecosystem have improved the accessibility and utility of BEL and have motivated its wider adoption (34). Unfortunately, unlike many of the other focuses of web applications, algorithms have remained confined to use by bioinformatics and inaccessible to a wider audience of researchers across disciplines. Last, but not least, the ecosystem of BEL-specific web applications is small and does not include a service for parsing, validating, compiling and converting BEL.

There are still several unmet needs for users that motivate the development of new web applications. Generally, there is still the need to enable complex exploration and visualization as well as to make algorithms and analyses generally accessible and reusable. Specifically to BEL, there is a need to make parsing, validating, compiling and converting facile and user-friendly. Finally, an integrative knowledge discovery environment that comprises many of the previously mentioned features would be greatly beneficial to the BEL and overarching scientific community. Here, we present BEL Commons, a web application that addresses these unmet needs and is a first attempt at building such an environment.

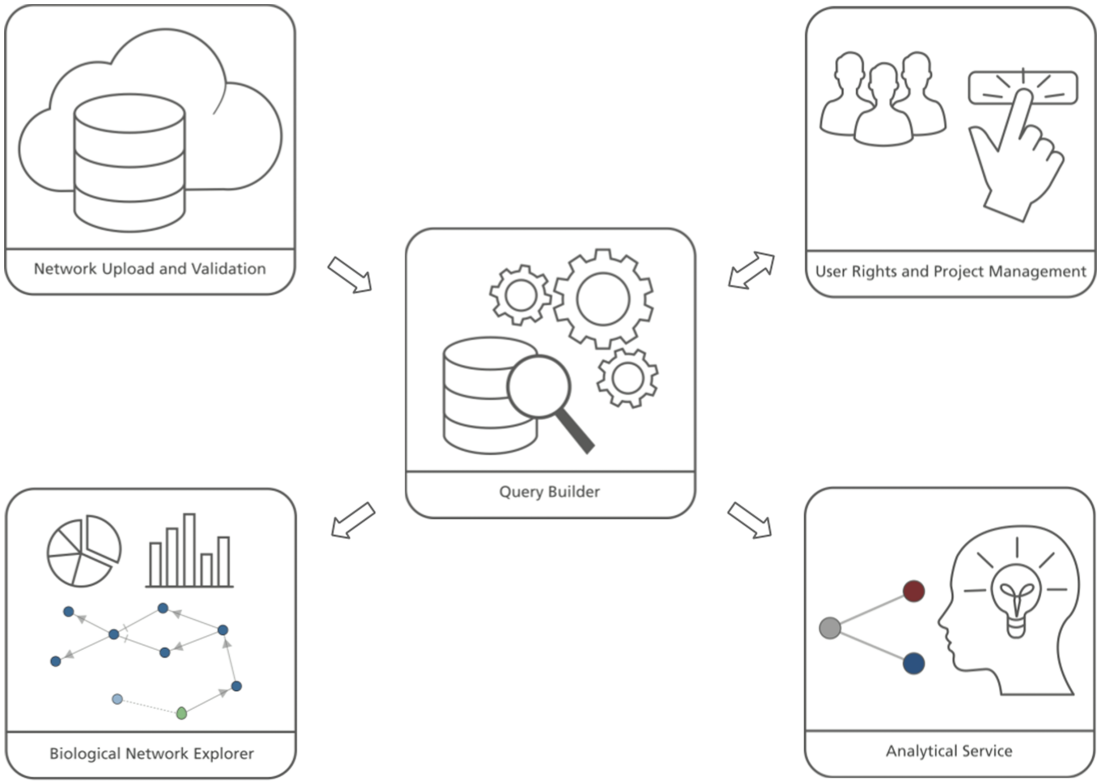


Figure 1. BEL Commons comprises several components: (i) the network uploader and validator, (ii) user rights and project management, (iii) the query builder, (iv) the biological network explorer and (iv) the analytical service.

Implementation and components

The user interface of BEL Commons integrates several features from the variety of previously mentioned web applications for systems and networks biology (Figure 1). It contains five main components: (i) the network uploader, where users can upload, parse, validate and compile BEL as well as generate several summaries of its contents; (ii) user rights and project management, where users can share their networks with various granularities; (iii) the query builder, where users can interactively query networks and make transformations; (iv) the biological network explorer, where users can visualize, explore and further modify networks; and (v) the analytical service, where users can run heat diffusion experiments with *-omics* data. Below, we elaborate on their functionalities and typical use cases. Implementation details can be found in the Supplementary Data.

Network uploader

The first point of entry for many users of BEL Commons will be through its BEL uploader, which allows users to choose a file from their computer to upload and to toggle common parsing and compilation parameters. After submitting, users' files are sent to an asynchronous task queue, implemented with RabbitMQ (https://www.rabbitmq.com) and Celery (http://www.celeryproject.org), which performs parsing, validation and compilation with PyBEL (34) in the background. Errors and warnings encountered during parsing are enumerated, statistics over the resulting network are produced, biological network motifs are identified and, finally, the submitter is notified upon completion.

The parsing errors and warnings are categorized first as syntactic or semantic then with much more detail, as described on the BEL Commons help page (https://bel-commons.scai.fraunhofer.de/help/parser). Each is presented with provenance information including the line, line number and position so curators can quickly make changes. Recurring errors and warnings are identified and grouped separately to allow curators to quickly make impactful improvements. Finally, a faceted search is presented for situations where an overwhelming number of errors and warnings are present.

The statistical summary (Figure 2A) of the network presents information about the contents of the network and also network theoretic measurements of the full network. Several charts are generated depicting the types and number of nodes, edges, modifications, namespaces, annotations and citations existing in the network. Furthermore, scalar

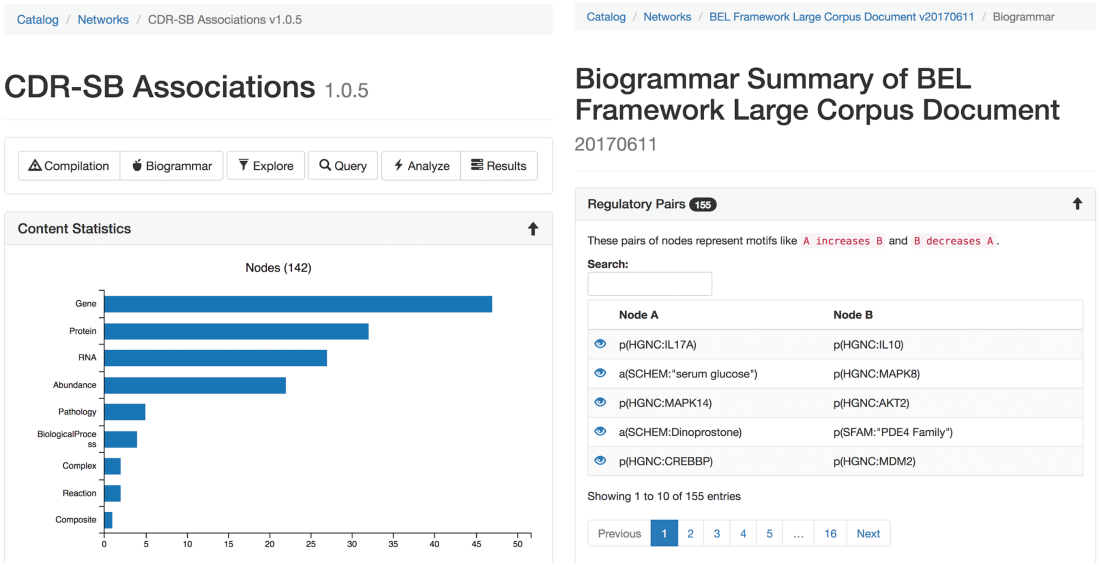


Figure 2. The statistical (**A**, left) and biogrammar (**B**, right) summary pages.

values describing network properties such as network density, average node degree and node overlap with other networks in BEL Commons using the Szymkiewicz–Simpson coefficient are listed.

The 'biogrammar' summary (Figure 2B) of the network presents an analysis of network motifs generalized from the analysis of transcriptional network motifs presented by Alon (35) for use with knowledge networks. BEL Commons focuses mainly on simple motifs that are informative to the robustness, correctness and applicability of a given knowledge work. The most simple motif is the contradictory pair, where knowledge has been curated stating *A increases B* but also *A decreases B*. Searching this motif in NeuroMMSig identified the contradiction that RB1 has been shown to not only increase the transcriptional activity of E2F4 by Li *et al.* (36) but also decrease the transcriptional activity of E2F4 (37). Another motif is an inconsistent negative correlation triple, where knowledge has been curated stating *A negatively correlates with B*, *B negatively correlates with C* and *C negatively correlates with A*. After, several factors can be used to estimate the confidence of the correctness and applicability of the statements such as biological context (e.g. cell line, tissue, disease), reference type (e.g. experimental paper, review paper, database), text location (e.g. abstract, introduction, methods, discussion) or the date of publication. BEL Commons currently only identifies predefined, small motifs containing two or three nodes, but could be extended to automatically find larger ones with the caveat that their biological meanings are more difficult to interpret.

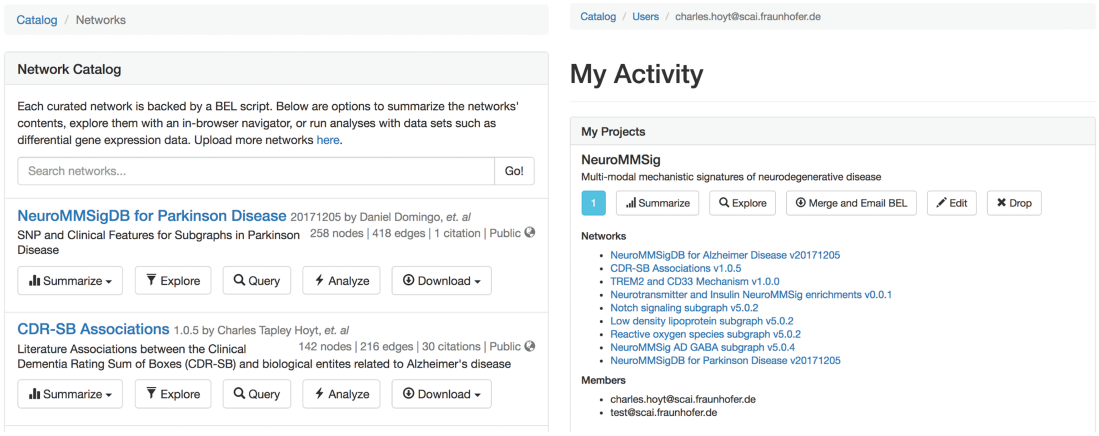


Figure 3. The network catalog (**A**, left) and user activity page (**B**, right).

Table 1. Statistics over a selection of the resources publicly initially available in BEL Commons. These numbers are accurate to the best of our ability, but may not reflect nominal values from their sources depending on the ability of PyBEL to parse their contents

Resource	Networks	Nodes	Edges	Citations
Selventa Example Corpora (http://resources.openbel.org)	5	16 339	36 971	5083
Causal Biological Networks Database (38)	138	5343	28 766	4580
NeuroMMSig (31)	8	1411	3221	201

User rights managements and collaboration

Upon BEL upload, users are presented with the option to make the resulting network either public or private. Networks can be uploaded privately for use during research then later released publicly to accompany a publication and share their work with the scientific community. The network catalog (Figure 3A) dynamically shows users only networks for which they have the appropriate permissions.

Users can create projects that allow for multiple users to mutually share networks. For example, a curation project in a given disease area could contain networks generated from the efforts of multiple curators. Projects can generate a merged network that can be summarized, explored, analyzed and exported with the same tools available for standalone networks. Users can access their private activity page (Figure 3B), which provides a global summary over their projects, networks, queries, data sets and experiments.

We do not presume all users plan to produce their own BEL, especially with the growing number of both general and context-specific publicly available resources. In light of this, we have included several of these resources in BEL Commons for these users (Table 1). The catalog of networks for which users have the appropriate permissions can be accessed directly from the home page of BEL Commons.

Query builder

While networks arising from BEL can be readily merged, it becomes increasingly difficult to visualize and explore large networks and combinations of networks. The query builder assists users in generating powerful, precise and expressive queries that find the most relevant and interesting subnetworks in three steps: (i) users search and select relevant networks; (ii) users generate a subnetwork specifying the most interesting nodes, edge annotations and references with their preferred *seeding method(s)*; and (iii) users select transformations (e.g. enrichments, selections, filters) to apply to the resulting subnetwork.

The first step in the query builder allows users to select networks relevant to their scientific questions. While it still remains computationally feasible to query over a merged view of the entire catalog of networks, users have the opportunity to pre-select the most relevant networks on the basis of their specificities toward target disease areas, curation methods or any other appropriate criteria described in their metadata. Additionally, other high-quality structured knowledge resources such as protein families (39, 40), biochemical reactions (41, 42) and gene orthologs [e.g. Entrez Gene (43), Mouse Genome Informatics (MGI) (44), Rat Genome Database (RGD) (45) etc.] can be included to enrich networks from curated BEL. Later, we will show how this novel feature can be used to enrich networks as a preprocessing step to connect disparate components before analysis.

The second step in the query builder allows users to generate subnetworks based on nodes, edge annotations and references of interest by using one or several 'seeding methods'. An example of seeding method that most network-related web applications implement is the retrieval of the Nth neighbors of a given node or set of nodes. Because neighborhood queries are often insufficient to capture complex biology, BEL Commons implements several additional seeding methods, enumerated in Table 2, that allow users to take advantage of the directionality, polarity and rich annotations inherent to networks from BEL. Using these seeding methods, the query builder allows scientists to ask scientific questions like the one proposed in the following scenario: the leukemia drug, nilotinib, triggers cells to remove faulty components, including ones associated with several brain diseases (46). In 2015, the Georgetown University Medical Center published findings that the drug had a therapeutic effect on patients with Alzheimer's and Parkinson's diseases (47). Though the drug's mechanism of action is currently unknown, a path search between nilotinib and these diseases suggests it could be by decreasing phosphorylation of the Tau protein, which may have a therapeutic effect in both disease contexts, through inhibition of ABL1 (48, 49).

The third step in the query builder allows users to specify transformations (e.g. enrichments, selections, filters) to apply to network resulting from the assembly in the first step then the seeding in the second step. Users may select basic transformations, such as deleting single nodes or edges, to more complex transformations such

Table 2. Seed methods available in the query builder

Seed method	Description
Nth neighbors	This induces a subnetwork over nodes in paths of length less than or equal to N from any query node, terminating at any node, including ones not included in the query.
Upstream subnetwork	This induces a subnetwork over nodes with causal edges targeting the query nodes then repeats a second time for that subnetwork in order to include a second layer. Finally, induces all causal edges between resulting nodes.
Downstream subnetwork	This induces a subnetwork over nodes with causal edges originating from the query nodes then repeats a second time for that subnetwork in order to include a second layer. Finally, induces all causal edges between resulting nodes.
Shortest paths	This induces a subnetwork over all nodes in the shortest paths between any pair of query nodes, implemented by NetworkX (50).
All paths	This induces a subnetwork over all nodes in all of the paths (length less than seven) between any pair of query nodes, implemented by NetworkX (50).
Provenance	This builds a subnetwork from all edges with provenance from articles with the given PubMed identifiers.
Authors	This builds a subnetwork from all edges from articles written by authors with the given names.
Annotations	This builds a subnetwork from all edges matching given biological or contextual annotations.

as selecting a subnetwork only consisting of causal edges. BEL Commons dynamically loads transformation functions from PyBEL such that new functions, pipelines and workflows for processing networks can be written quickly and made available to users. A full list is available at https://bel-commons.scai.fraunhofer.de/help/query-builder.

Each query is saved with a unique identifier such that queries can be rerun, shared, merged and compared. Rather than storing the results of queries, the selection, seeding and transformations are stored as a 'transaction' so that they can be applied to new assemblies, for example, when a network is updated. Effectively, queries correspond to an experimental protocol for processing raw networks before visualization, exploration, analysis and interpretation. However, the construction of a query is not the end of its life. The next section summarizing the biological network explorer describes how queries can be extended and evolve during the process of scientific inquiry.

Biological network explorer

The biological network explorer provides users with easy ways to visualize BEL networks to interpret their underlying structures, to investigate the metadata on nodes and edges and to interactively update networks as they explore (Figure 4). It is built with D3.js (https://d3js.org) to render networks with a force-directed layout algorithm that can be panned and zoomed. Because the complexity of biological networks often limits the utility of automated layouts (51), users can also manually drag and reposition nodes. Furthermore, users can adjust the edge length parameter of the algorithm to rarefy densely grouped nodes and improve readability. The networks are styled with minimum visual clutter and make use of easily distinguishable colors rather than obtrusive shapes for nodes as well as patterns and colors for different types of edges.

The explorer has several contextual actions registered to the nodes and edges. Users can left-click a node to populate the information box located below the explorer with information from external data sources [e.g. Entrez Gene, ChEBI (52), ExPASy (40), Gene Ontology (53)] gathered from Bio2BEL services (https://github.com/bio2bel) and the EMBL Ontology Lookup Service (54). Alternatively, users can right-click a node to open a contextual menu that enables further modification to the network (e.g. delete the node, add the neighbors of the node to the network) that are interactively appended to the original query used to render the visualization. The contents of the network can also be further modified by the inline query builder, which allows additional transformations to be applied interactively. The query history is displayed at the bottom of the explorer, new changes are highlighted in red, and because queries are stored as transactions, changes can be reverted with an 'undo' button.

When an edge is clicked, the information box is populated with relevant citations, evidences and annotations. Each edge is linked to a voting and commenting system so domain-specific experts, curators and bioinformaticians can engage in discussion on the correctness and robustness of the chosen representation of knowledge.

To the right of the explorer is the filter tool box, which incorporates a novel approach to filtering and exploring networks using a linked hierarchical explorer of the terminologies/ontologies annotated to the edges in the currently displayed networks. Users can search and select groups of annotations to filter the network. For example, this could be useful to exclude edges asserted from research on cell lines that are not relevant. The filter tool box has three additional

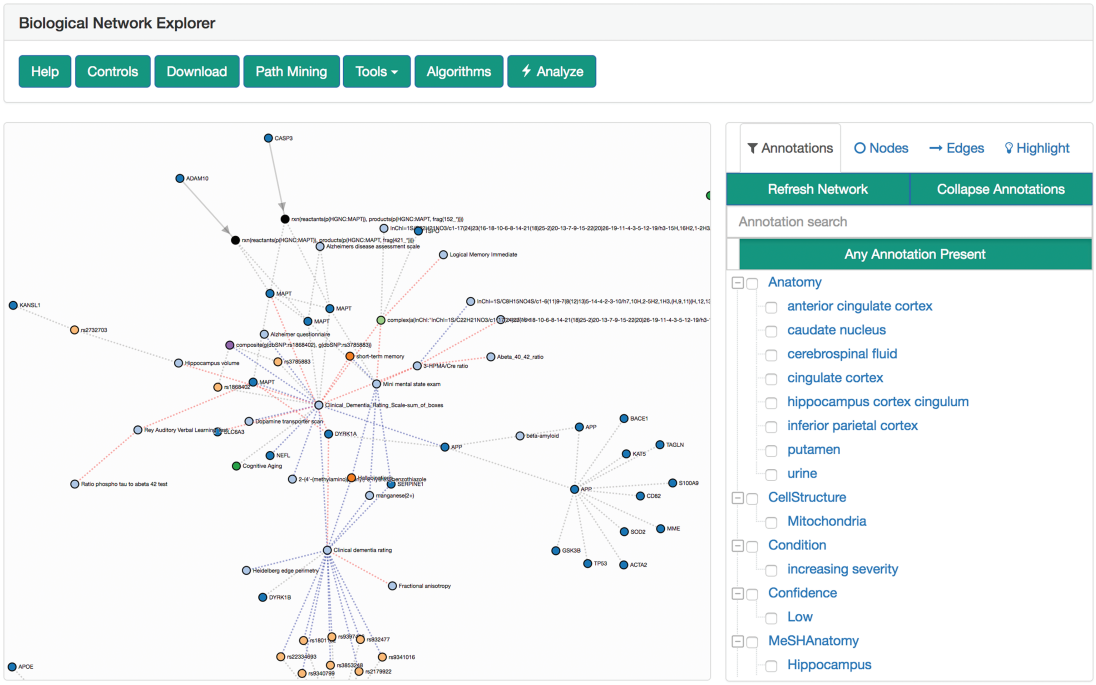


Figure 4. The biological network explorer and related navigation components. Truncated from this image are the node information and query information boxes.

tabs: nodes, edges and highlight. Users can either search for specific nodes and edges in their corresponding tabs or use the highlight tab to select nodes and edges with specific properties to highlight in the network.

Above the explorer is the general tool box that includes several additional interactions for exploration, analysis and export of the network using the serializers described by Hoyt *et al.* (34). Notably, it contains a path mining tool that enables path searches between given nodes with fine granular, configurable settings (e.g. directedness, path search algorithm, application of filters for pathologies etc.). Hence, it can immediately be used to identify the causal root affecting two nodes or generate hypothetical links across modes and scales.

The visualization can be further modified by resizing the nodes corresponding to the results of topological or data-driven analyses, such as their degree, betweenness centrality or by the results of an experiment (e.g. heat diffusion with *-omics* data, see next section) in order to identify novel biological entities.

Finally, there are several alternatives to exploring networks that are too large to render in-browser due to the limits of JavaScript-based graphics. First, the network catalog opts to present users with a random subsampling of large networks. If a network in the explorer becomes too big, then the explorer prompts the usage of the filter tool box to identify a more relevant, smaller network. Otherwise, users can export the current network to multiple formats for use in desktop visualization applications.

Analytical service

Khatri *et al.* categorized algorithms for analyzing pathways and networks in three types: over-representation analysis, functional class scoring and pathway topology (26). Algorithms of each type have been developed for a wide variety of applications, data formats and network types, but most are difficult to use and few are specific to networks from BEL. The BEL Commons analytical service begins to address this issue by coupling a heat diffusion workflow to the query builder and explorer to create a more seamless user experience.

In the context of network science, heat diffusion refers to annotating a scalar value to each node (i.e. heat) and simulating how it spreads through nodes' adjacent edges to their neighbors over several iterations. It has been used successfully in systems and networks biology to assess the connectivity of nodes and identify important subnetworks as demonstrated by Leiserson *et al.* with the HotNet2 algorithm (55).

BEL Commons exposes a similar workflow, previously presented by Hoyt *et al.* (34), that has the added behavior based on the polarity of causal edges—when heat crosses a *decreases* edge, its sign is flipped to better capture the

aggregate effect of heat flowing from several edges with mixed polarity to a single node. Because BEL contains several entity types, users are presented with the final heats on biological process nodes to assist in interpreting which processes are dysregulated in the experiment. A more detailed description of this method can be found at https://bel-commons.scai.fraunhofer.de/help/heat-diffusion.

Users can upload pre-processed, high throughput *-omics* experiments (e.g. differential gene expression data) and map them to a network either by starting in the network catalog or through the biological network explorer's toolbox. The network and *-omics* data are then sent to the task queue to perform the heat diffusion workflow. Upon completion, users are notified via email with a link to the results page that shows statistics and data visualization. Finally, users are also able to overlay the results on the original network in the biological network explorer.

In the following section, the query builder, biological network explorer and analytical service are used to assess several differential gene expression experiments representing patients with Alzheimer's disease at different disease states using NeuroMMSig networks.

Application scenario

This section describes a use case in which a disease-specific network for Alzheimer's disease is assembled and pre-processed. First, the network is explored with the biological network explorer; second, it is enriched; and finally, it is analyzed with differential gene expression data in order to identify patterns of dysregulation of biological processes that are specific to disease progression stages.

We uploaded several BEL documents describing Alzheimer's disease pathophysiology generated during the AETIONOMY project (https://www.aetionomy.eu) that were originally stored in NeuroMMSig (https://neurommsig.scai.fraunhofer.de) to BEL Commons. We began by using the query builder to search for and select all of these networks. Next, we used the 'Annotations' seeding method (Table 2) to generate a subnetwork composed of edges that had been annotated in BEL with membership in the following candidate mechanisms of Alzheimer's disease pathophysiology from NeuroMMSig. We chose the low density lipoprotein subgraph, the GABA subgraph, the notch signaling subgraph and the reactive oxygen species subgraph because they are dysregulated at different disease stages. Later, we will capture these different progression patterns by running the heat diffusion workflow with stage-specific differential gene expression data.

We used the query builder to apply several transformations to pre-process the network, including (i) enrichment of the network with the members of all protein families and protein complexes; (ii) deletion of nodes having the MGI and RGD namespaces that respectively correspond to genes from mice and rats; (iii) removing pathology nodes, which often are uninformative hubs in disease-specific networks; and (iv) extracting only causal edges. Finally, we submitted the query and visualized the network.

The biological network explorer showed that there were several disconnected components in the resulting network. Further, there were several cases where the gene, RNA or protein, such as the GABRA4 gene and RNA, were in different components. We used the tool box above the explorer to apply an additional filter, 'Enrich Protein And RNA Origins', which added the corresponding RNA for each protein then the corresponding gene for each ribonucleic acid/micro-ribonucleic acid (RNA/miRNA) in the network. Finally, we applied 'Collapse Variants' and 'Collapse to Genes' to simplify the network by collapsing all corresponding genes, RNAs, proteins and their variants to a single node for use with the heat diffusion workflow.

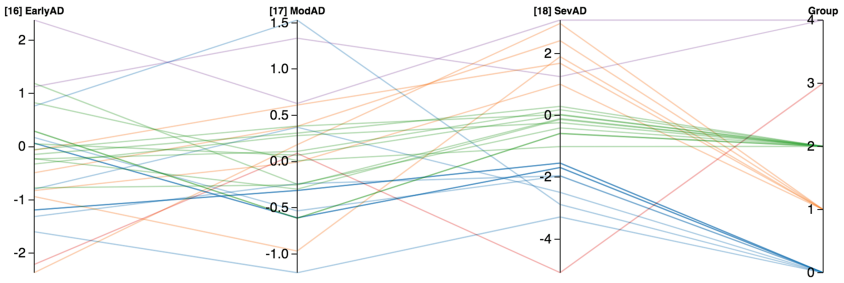
After clicking the 'Analyze' button above the explorer, we uploaded three differential gene expression analyses corresponding to patients with Alzheimer's disease in three disease stages (i.e. early, moderate and severe) from Blalock *et al.* [GSE28146, (56)] pre-processed with GEO2R (57). We applied the heat diffusion workflow to the previously generated network using each of the three differential gene expression analyses in parallel and displayed the results (i.e. the final heat on each biological process in the network) together with the parallel coordinate display in BEL Commons (Figure 5). Because each biological process has a final heat corresponding to experiments for the three disease stages, the parallel plot directly allows interpretation of progression patterns. We used BEL Commons to apply a K-Means clustering with $K = 5$ to assist in identifying clusters of biological processes with similar progression patterns and color them accordingly. In Figure 5, biological processes in *group 0* (blue) tended to decrease only at severe onset of disease (e.g. glial cell differentiation). Conversely, biological processes in *group 1* (orange) tended to increase throughout progression of disease (e.g. Notch signaling pathway). Finally, *group 2* (green) processes remained relatively unchanged (e.g. lipid metabolic process) through the progression of disease and *group 4* (purple) remained consistently elevated (e.g. apoptotic process). The complete results of this experiment as well as a tutorial for reproduction can be found in the Supplementary Data.

While BEL has inherent limits in its temporal expressivity, using temporal data in analysis is an initial attempt to overcome these limits. Complex diseases like Alzheimer's

Summary of Experiments

Identifier	Query	Permutations	Omic	Name	Description	
16	14	250	13	EarlyAD.csv	Patients from GSE28146 with early Alzheimer's disease	Delete
17	14	250	14	ModAD.csv	Patients from GSE28146 with moderate Alzheimer's disease	Delete
18	14	250	15	SevAD.csv	Patients from GSE28146 with severe Alzheimer's disease	Delete

Comparison of Experiments Download Data as TSV



K-Means clustering with K=5 is enabled. Be aware that this is a randomized algorithm and results may not be consistent.

Type	Namespace	Name	[16] EarlyAD	[17] ModAD	[18] SevAD	Group
BiologicalProcess	GOBP	glial cell differentiation	-0.7989	0.371	-2.4916	0
BiologicalProcess	GOBP	Notch signaling pathway	-2.3671	0.1815	2.9763	1
BiologicalProcess	GOBP	lipid metabolic process	0.291	-0.6136	-0.5847	2
BiologicalProcess	GOBP	response to oxidative stress	-2.2128	0.088	-5.0908	3
BiologicalProcess	GOBP	apoptotic process	1.1241	1.3339	1.2571	4

Figure 5. A parallel coordinate plot of the final heats from biological processes from several Alzheimer's disease-specific networks after running the heat diffusion workflow with differential gene expression data from Blalock *et al*. (56) comparing patients at three stages of Alzheimer's disease.

disease must be studied with respect to its progression over time, and we believe that workflows like the one presented above could begin to provide insight to the genesis and progression of the disease in order to support patient stratification and precision medicine.

Discussion

No web application, however feature-rich, will ever satiate the desire and creativity of researchers to generate novel solutions to scientific problems. Even though BEL Commons has taken inspiration from many well-constructed services to build a knowledge discovery environment that enables researchers to explore knowledge and data in new ways, it still shares this limit. However, we are not discouraged, and we hope to make several improvements to BEL Commons in the future.

We would like to improve the interoperability of BEL Commons and the platform build on BEL itself by integrating open authentication systems like Open Researcher and Contributor Identifier (ORCID) (https://orcid.org) in order to harmonize identification of users across multiple web services and provide reliable provenance for networks, queries and analyses. We would also like to integrate further tools for converting BEL to Resource Description Framework (RDF) in order to connect BEL with other linked data. Further, we would like to improve exporters to other services, notably, NDEx, which have brighter outlooks on sharing and feedback systems. Recent developments in integrating Integrated Network and Dynamical Reasoning Assembler (INDRA) (58) with PyBEL enable conversion from BioPAX documents to BEL. A future update to BEL Commons will include an option to upload these documents as well.

We would like to integrate BEL Commons with other BEL-specific systems developed with different underlying technologies. First, integrating the BELIEF Dashboard to use the underlying network and edge store from PyBEL would enable a more thorough feedback and curation interface so users could not only vote on the correctness of edges, but also fix them directly. Second, the NeuroMMSig Mechanism Enrichment Server will be re-implemented completely with reusable PyBEL code and BEL Commons components in order to advance its goals of achieving patient stratification by using common algorithms and tools.

Conclusion

Along with recent improvements in generation of BEL content through text mining (INDRA) and serialization

of resources (Bio2BEL project, https://github.com/bio2bel), we believe that BEL Commons will make BEL more accessible to both academic and industrial users. We have made this application freely available at https://bel-commons.scai.fraunhofer.de.

Authors' contributions

C.T.H. and D.D.F. conceived the web application, implemented it and wrote this manuscript. M.H.A. reviewed the content.

Supplementary data

Supplementary data are available at *Database* Online.

Acknowledgements

We would like to thank all colleagues who assisted in testing and provided feedback in order to improve this work, especially to André Gemünd for his technical assistance. We would also like to thank Scott Colby for making our logos.

Funding

European Union / European Federation of Pharmaceutical Industries and Associations (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.

Conflict of interest. None declared.

References

1. Demir,E., Cary,M.P., Paley,S. *et al.* (2010) The BioPAX community standard for pathway data sharing. *Nat. Biotechnol.*, **28**, 1308.
2. Hucka,M., Bergmann,F.T., Hoops,S. *et al.* (2015) The Systems Biology Markup Language (SBML): language specification for level 3 version 1 core. *J. Integr. Bioinform.*, **12**, 382–549.
3. Le Novère,N., Hucka,M., Mi,H. *et al.* (2009) The systems biology graphical notation. *Nat. Biotechnol.*, **27**, 735–741.
4. Slater,T. (2014) Recent advances in modeling languages for pathway maps and computable biological networks. *Drug Discov. Today*, **19**, 193–198.
5. Saqi,M., Lysenko,A., Guo,Y.-K. *et al.* (2018) Navigating the disease landscape: knowledge representations for contextualizing molecular signatures. *Brief. Bioinformatics*, 1–15.
6. Valenzuela-Escárcega,M.A., Hahn-Powell,G., Hicks,T. *et al.* (2015) A domain-independent rule-based framework for event extraction. In: Proceedings of ACL-IJCNLP 2015 System Demonstrations, 127–132.
7. Allen,J.F., Swift,M. and De Beaumont,W. (2008) Deep semantic analysis of text. In: Proceedings of the 2008 Conference on Semantics in Text Processing STEP 08, **1**, 343–354.
8. Björne,J., Heimonen,J., Ginter,F. *et al.* (2011) Extracting contextualized complex biological events with rich graph-based feature sets. *Comput. Intell.*, **27**, 541–557.
9. Novichkova,S., Egorov,S. and Daraselia,N. (2003) MedScan, a natural language processing engine for MEDLINE abstracts. *Bioinformatics*, **19**, 1699–1706.
10. Madan,S., Hodapp,S., Senger,P. *et al.* (2016) The BEL information extraction workflow (BELIEF): evaluation in the BioCreative V BEL and IAT track. *Database*, **2016**, 1–17.
11. Ravikumar,K.E., Rastegar-Mojarad,M. and Liu,H. (2017) BELMiner: adapting a rule-based relation extraction system to extract biological expression language statements from bio-medical literature evidence sentences. *Database*, **2017**, 1–12.
12. Lai,P.T., Lo,Y.Y., Huang,M.S. *et al.* (2016) BelSmile: a biomedical semantic role labeling approach for extracting biological expression language from text. *Database*, **2016**, 1–9.
13. Rastegar-Mojarad,M., Komandur Elayavilli,R. and Liu,H. (2016) BELTracker: evidence sentence retrieval for BEL statements. *Database*, **2016**, 1–11.
14. Guryanova,S. and Guryanova,A. (2017) sbv IMPROVER: modern approach to systems biology. *Methods Mol. Biol.*, **1613**, 21–29.
15. Madan,S., Hodapp,S. and Fluck,J. (2015) BELIEF dashboard—a web-based curation interface to support generation of BEL networks. In: Proceedings of the Fifth BioCreative Challenge Evaluation Workshop, 409–417.
16. Slenter,D.N., Kutmon,M., Hanspers,K. *et al.* (2017) WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res.*, **46**, D661–D667.
17. Chou,C.H., Shrestha,S., Yang,C.D. *et al.* (2017) miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic Acids Res.*, **46**, D296–D302.
18. Davis,A.P., Grondin,C.J., Johnson,R.J. *et al.* (2016) The comparative toxicogenomics database: update 2017. *Nucleic Acids Res.*, **45**, D972–D978.
19. Kanehisa,M., Furumichi,M., Tanabe,M. *et al.* (2016) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.*, **45**, D353–D361.
20. Cerami,E.G., Gross,B.E., Demir,E. *et al.* (2010) Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res.*, **39**, D685–D690.
21. Pratt,D., Chen,J., Welker,D. *et al.* (2015) NDEx, the network data exchange. *Cell Syst.*, **1**, 302–305.
22. Bharadwaj,A., Singh,D.P., Ritz,A. *et al.* (2017) GraphSpace: stimulating interdisciplinary collaborations in network biology. *Bioinformatics*, **33**, 3134–3136.
23. Kutmon,M., van Iersel,M.P., Bohler,A. *et al.* (2015) PathVisio 3: an extendable pathway analysis toolbox. *PLoS Computational Biol.*, **11**, e1004085.
24. Shannon,P., Markiel,A., Owen,O. *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.
25. Franz,M., Lopes,C.T., Huck,G. *et al.* (2015) Cytoscape.js: a graph theory library for visualisation and analysis. *Bioinformatics*, **32**, 309–311.

26. Khatri,P., Sirota,M. and Butte,A.J. (2012) Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.*, **8**.
27. Merico,D., Isserlin,R., Stueker,O. *et al.* (2010) Enrichment Map: a Network-Based Method for Gene-Set Enrichment Visualization and Interpretation. *PLoS ONE*, **5**(11): e13984.
28. Cavalli,F.M.G., Remke,M., Rampasek,L. *et al.* (2017) Intertumoral heterogeneity within medulloblastoma subgroups. *Cancer Cell*, **31**, 737–754.e6.
29. Subramanian,A., Tamayo,P., Mootha,V.K. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. U S A*, **102**, 15545–15550.
30. Liberzon,A., Subramanian,A., Pinchback,R. *et al.* (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, **27**, 1739–1740.
31. Domingo-Fernández,D., Kodamullil,A.T., Iyappan,A. *et al.* (2017) Multimodal mechanistic signatures for neurodegenerative diseases (NeuroMMSig): a web server for mechanism enrichment. *Bioinformatics*, **33**, 3679–3681.
32. Catlett,N.L., Bargnesi,A.J., Ungerer,S. *et al.* (2013) Reverse causal reasoning: applying qualitative causal knowledge to the interpretation of high-throughput data. *BMC Bioinformatics*, **14**, 340.
33. Martin,F., Thomson,T.M., Sewer,A. *et al.* (2012) Assessment of network perturbation amplitudes by applying high-throughput data to causal biological networks. *BMC Syst. Biol.*, **6**, 54.
34. Hoyt,C.T., Konotopez,A. and Ebeling,C. (2018) PyBEL: a computational framework for Biological Expression Language. *Bioinformatics*, **34**, 703–704.
35. Alon,U. (2007) Network motifs: theory and experimental approaches. *Nat. Rev. Genet.*, **8**, 450–461.
36. Li,J.M., Hu,P.P., Shen,X. *et al.* (1997) E2F4-RB and E2F4-p107 complexes suppress gene expression by transforming growth factor beta through E2F binding sites. *Proc. Natl. Acad. Sci. U S A*, **94**, 4948–4953.
37. Kohn,K.W. (1999) Molecular interaction map of the mammalian cell cycle control and DNA repair systems. *Mol. Biol. Cell*, **10**, 2703–2734.
38. Boué,S., Talikka,M., Westra,J. *et al.* (2015) Causal biological network database: a comprehensive platform of causal biological network models focused on the pulmonary and vascular systems. *Database*, **2015**, bav030.
39. Finn,R.D., Attwood,T.K., Babbitt,P.C. *et al.* (2016) InterPro in 2017—beyond protein family and domain annotations. *Nucleic Acids Res.*, **45**, D190–D199.
40. Gasteiger,E., Gattiker,A., Hoogland,C. *et al.* (2003) ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.*, **31**, 3784–3788.
41. Placzek,S., Schomburg,I., Chang,A. *et al.* (2017) BRENDA in 2017: new perspectives and new tools in BRENDA. *Nucleic Acids Res*, **45**, D380–D388.
42. Morgat,A., Lombardot,T., Axelsen,K.B. *et al.* (2017) Updates in Rhea—an expert curated resource of biochemical reactions. *Nucleic Acids Res.*, **45**, D415–D418.
43. Maglott,D., Ostell,J., Pruitt,K.D. *et al.* (2011) Entrez gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **39**, 52–57.
44. Blake,J.A., Eppig,J.T., Kadin,J.A. *et al.* (2017) Mouse Genome Database (MGD)-2017: community knowledge resource for the laboratory mouse. *Nucleic Acids Res.*, **45**, D723–D729.
45. Shimoyama,M., De Pons,J., Hayman,G.T. *et al.* (2015) The Rat Genome Database 2015: genomic, phenotypic and environmental variations and disease. *Nucleic Acids Res.*, **43**, D743–D750.
46. Hebron,M.L., Lonskaya,I. and Moussa,C.E. (2013) Nilotinib reverses loss of dopamine neurons and improves motor behavior via autophagic degradation of alpha-synuclein in Parkinson's disease models. *Hum. Mol. Genet.*, **22**, 3315–3328.
47. Pagan,F., Hebron,M., Valadez,E.H. *et al.* (2016) Nilotinib effects in Parkinson's disease and dementia with Lewy bodies. *J. Parkinsons Dis.*, **6**, 503–517.
48. Duveau,D.Y., Hu,X., Walsh,M.J. *et al.* (2013) Synthesis and biological evaluation of analogues of the kinase inhibitor nilotinib as Abl and Kit inhibitors. *Bioorg. Med. Chem. Lett.*, **23**, 682–686.
49. Derkinderen,P., Scales,T.M.E., Hanger,D.P. *et al.* (2005) Tyrosine 394 is phosphorylated in Alzheimer's paired helical filament tau and in fetal tau with c-Abl as the candidate tyrosine kinase. *J. Neurosci.*, **25**, 6584–6593.
50. Hagberg,A.A., Schult,D.A. and Swart,P.J. (2008) Exploring network structure, dynamics, and function using NetworkX. In: Proceedings of the 7th Python in Science Conference (SciPy 2008), (SciPy), 11–15.
51. Schreiber,F., Dwyer,T., Marriott,K. *et al.* (2009) A generic algorithm for layout of biological networks. *BMC Bioinformatics*, **10**, 1–12.
52. Hastings,J., De Matos,P., Dekker,A. *et al.* (2013) The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res.*, **41**, 456–463.
53. Carbon,S., Dietze,H., Lewis,S.E. *et al.* (2017) Expansion of the gene ontology knowledgebase and resources: the gene ontology consortium. *Nucleic Acids Res.*, **45**, D331–D338.
54. Cote,R., Jones,P., Apweiler,R. *et al.* (2006) The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries. *BMC Bioinformatics*, **7**, 1–7.
55. Leiserson,M.D.M., Vandin,F., Wu,H.-T. *et al.* (2015) Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.*, **47**, 106–114.
56. Blalock,E.M., Buechel,H.M., Popovic,J. *et al.* (2011) Microarray analyses of laser-captured hippocampus reveal distinct gray and white matter signatures associated with incipient Alzheimer's disease. *Journal of Chemical Neuroanatomy.* **37**, 62–70.
57. Barrett,T., Wilhite,S.E., Ledoux,P. *et al.* (2013) NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.*, **41**, D991–D995.
58. Gyori,B.M., Bachman,J.A., Subramanian,K. *et al.* (2017) From word models to executable models of signaling networks using automated assembly. *Mol. Syst. Biol.*, **13**, 954.

Conclusions

4 Re-curation and rational enrichment of knowledge graphs in Biological Expression Language

Introduction

Re-curation and Rational Enrichment of Knowledge Graphs in Biological Expression Language

Charles Tapley Hoyt[1,2,*], Daniel Domingo-Fernández[1,2], Rana Aldisi[1,2], Lingling Xu[1,2], Kristian Kolpeja[1], Sandra Spalek[1], Esther Wollert[1], John Bachman[3], Benjamin M. Gyori[3], Patrick Greene[3], and Martin Hofmann-Apitius[1,2]

1. Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Sankt Augustin 53754, Germany
2. Bonn-Aachen International Center for IT, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn 53115, Germany
3. Laboratory of Systems Pharmacology, Harvard Medical School, 200 Longwood Ave, 02115 Boston, MA, USA

***Corresponding Author**: Hoyt, C. T., Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Sankt Augustin 53754, Germany. Telephone details; +49 2241 14-2268

Keywords: Natural language processing, Information extraction, Biocuration, Biological Expression Language, Knowledge graphs

Abstract

The rapid accumulation of new biomedical literature not only causes curated knowledge graphs to become outdated and incomplete, but also makes manual curation an impractical and unsustainable solution. Automated or semi-automated workflows are necessary to assist in prioritizing and curating the literature to update and enrich knowledge graphs.

We have developed two workflows: one for re-curating a given knowledge graph to assure its syntactic and semantic quality and another for rationally enriching it by manually revising automatically extracted relations for nodes with low information density. We applied these workflows to the knowledge graphs encoded in Biological Expression Language from the NeuroMMSig database using content that was pre-extracted from MEDLINE abstracts and PubMed Central full text articles using text mining output integrated by INDRA. We have made this workflow freely available at https://github.com/bel-enrichment/bel-enrichment.

Database URL: https://github.com/bel-enrichment/results

Background

The rapid accumulation of unstructured knowledge in the biomedical literature has motivated its structuring and formalization so computers can assist in large-scale reasoning and interpretation. Several standard formats have been proposed for storing newly structured knowledge, including Systems Biology Markup Language (SBML; Hucka *et al.*, 2003), Biological Pathways Exchange Language (BioPAX; Demir *et al.*, 2010), Biological Expression Language (BEL; Slater, 2014), Gene Ontology Causal Assembly Models (CAMs; Carbon *et al.*, 2017). Accompanying these standards are public repositories containing content generated both in academic and industrial contexts such as the BioModels Database (Glont *et al.*, 2018), Pathway Commons (Cerami *et al.*, 2011), NDEx (Pratt *et al.*, 2015), Bio2RDF (Belleau *et al.*, 2008), Open PHACTS (Williams *et al.*, 2012), and BEL Commons (Hoyt *et al.*, 2018). Additionally, a significant number of databases use custom formats for knowledge that are not appropriate for formalization in a standard format.

Even though each standard focuses on different aspects of modeling knowledge in systems and networks biology, they all give rise to knowledge graphs (KGs) consisting of biological entities (nodes), their interrelations (edges), and their associated metadata. While KGs have been useful for qualitative modeling of biochemical networks (Rausanu *et al.*, 2015; Yugi *et al.*, 2016), cellular signaling (Pilalis *et al.*, 2015; Pon *et al.*, 2015; Tripathi *et al.*, 2015), gene regulatory pathways and genetic interactions (Kandasamy *et al.*, 2010; Kamburov *et al.*, 2013), metabolic pathways (Caspi *et al.*, 2016; Wishart *et al.*, 2018), and other systems biology applications, there are several challenges associated with their use. First, they contain noise arising from curation, from the loss of information due to representation, and from normalization of different knowledge representations (Nickel *et al.*, 2016; Mihindukulasooriya *et al.*, 2017; Pujara *et al.*, 2017). Second, they are generally an incomplete representation of the current state of scientific knowledge due to the large amount of uncurated, unstructured knowledge in the literature. Third, they progressively become out-of-date as scientific experimentation and investigation elucidates new knowledge (Wadi *et al.*, 2016). Finally, they often lack biological contextual information such as organelle-, cell-, cell line-, tissue-, organ-, phenotype-, or disease-specificity (Hofmann-Apitius *et al.*, 2015; Saqi *et al.*, 2018).

KGs also suffer from issues in the normalization and mapping of entities. Though interoperability standards and resources like the Minimal Information Required in the Annotation of Models (MIRIAM;

Laibe *et al.*, 2007) and Identifiers.org (Juty *et al.*, 2012) have been developed and implemented to promote the semantic interoperability of biological models (and by extension, KGs), curators often encounter concepts that are not present in high-quality, publicly available terminologies and can not capture the incident knowledge in a semantically meaningful way. These situations require enriching previously existing terminologies or, in some cases, developing new ones. For situations when the appropriate concept/term is unclear, several tools have been developed and made freely available to the community to help curators build semantically interoperable models including the Ontology Lookup Service (OLS; Cote *et al.*, 2007), the Ontology Mapping Service (OxO; https://www.ebi.ac.uk/spot/oxo), Zooma (https://www.ebi.ac.uk/spot/zooma), and CEDAR Workbench (Gonçalves *et al.*, 2017). Further, recent work from Domingo-Fernández *et al.* on mapping pathways between major databases (Domingo-Fernández *et al.,* 2018) and a critical assessment of their overlaps and contradictions (Domingo-Fernández *et al.,* 2019) has shown that that the adoption of standards like MIRIAM has been slow and that while the syntax of the varying formats used by each database may be correct, their semantic interoperability is still lacking.

Motivation

Accurately structuring and formalizing the unstructured knowledge in the biomedical literature requires careful planning and manual effort from trained curators. The scope of a given project must be defined based on its scientific goals (e.g., to support the interpretation of data, to generate a disease-specific knowledgebase, etc.) and limited in its literature content sources (e.g., abstracts, full text, patents, etc.) based on a project-specific metric for quality and relevance — both of which are nebulous in description and difficult to generate. The scope must also be limited to certain classes of biological entities, their interrelations, and the standard formats that are capable of expressing them. For instance, the entities, relations, and formats used during curation are different for protein complex assemblies curated by the Complex Portal (Meldal *et al.*, 2015) and regulatory interactions curated by the Signaling Network Open Resource (SIGNOR; Perfetto *et al.*, 2016). Similarly, curation guidelines must be defined reflecting these limits. For example, the guidelines of a project designed to model Tau aggregation inhibitors from the chemistry literature might encourage the curators to include direct binding partners of those inhibitors (e.g., GSK-3β, CDK5, etc.) but explicitly exclude the biological mechanisms through which the inhibitors' targets result in Tau aggregation that would better be curated during a different project focusing on capturing molecular biology from its primary literature. While there is no alternative to proper planning, several semi-automated curation workflows such as BELIEF (Madan *et al.*, 2016) and

the sbv IMPROVER (Guryanova *et al.*, 2017) provide assistance by automatically detecting entities and relations for curators to accept or fix in order to increase productivity and enforce correct syntax and semantics. However, these and similar systems are limited in their ability to capture the relevant chemistry and biology, and reversion to manual curation is often necessary. Finally, the issues of insufficient resources and fixed timelines apply to most curation projects, as aptly described by Rodríguez-Esteban (2015).

In the AETIONOMY project (https://www.aetionomy.eu), we manually curated NeuroMMSig, an inventory of multiscale and multimodal knowledge graphs that capture mechanistic knowledge in the context of neurological disorders (Domingo-Fernández *et al.*, 2017). We encoded it in BEL because it is appropriate for qualitative causal, correlative, and associative relationships between biological entities, processes, and measurements across modes and scales. However, it is currently suffering from the issues we have previously described: it has not been assessed for confidence, is becoming outdated, and needs to be enriched following a rational approach that best prioritizes the flood of recent literature.

To address this, we have developed and applied two workflows, described in this paper: the first is for re-curating existing BEL documents to ensure their syntactic and semantic correctness in a scenario where there was neither prior syntax validation, curation guidelines for entity nomenclature, nor a second curator for achieving inter-annotator agreement. The second is a semi-automated algorithm and reproducible workflow for updating and rationally enriching an existing KG that lessens the burden of identifying relevant literature, reduces the overhead, as defined by Rodríguez-Esteban, and generates more, higher quality, relevant content.

We applied these workflows to a selection of knowledge graphs in NeuroMMSig and evaluated the curation effort (time) and quality in comparison to purely manual curation and other previously reported semi-automated curation workflows. We increased the number of nodes and edges in the selected knowledge graphs respectively by approximately five and seven times while maintaining the specificity of the knowledge graphs. With an improvement to the content underlying NeuroMMSig, the mechanism enrichment algorithm on its corresponding web service can return more correct and robust results to support the analysis of neuroimaging and genomics data for clinical trials in Alzheimer's disease, Parkinson's disease, and epilepsy. Finally, we have made this workflow freely available at https://github.com/bel-enrichment/bel-enrichment so others can include it in their own curation workflows.

Methods

We first present the re-curation workflow for syntactic and semantic quality assurance before presenting our proposed approach for updating and rational enrichment.

Syntactic Quality Assurance

We developed a workflow using git (https://git-scm.com), GitHub (https://github.com), PyBEL (Hoyt *et al.*, 2017), and a novel PyBEL extension PyBEL-Git (Hoyt, 2018) in order to identify and address syntactical issues in the BEL documents generated during the AETIONOMY project (https://www.aetionomy.eu; Irin *et al.*, 2015; Kodamullil *et al.*, 2015; Naz *et al.*, 2016; Emon *et al.*, 2017; Hoyt and Domingo-Fernández *et al.*, 2018) and exposed through the NeuroMMSig mechanism enrichment server (Domingo-Fernández *et al.*, 2017).

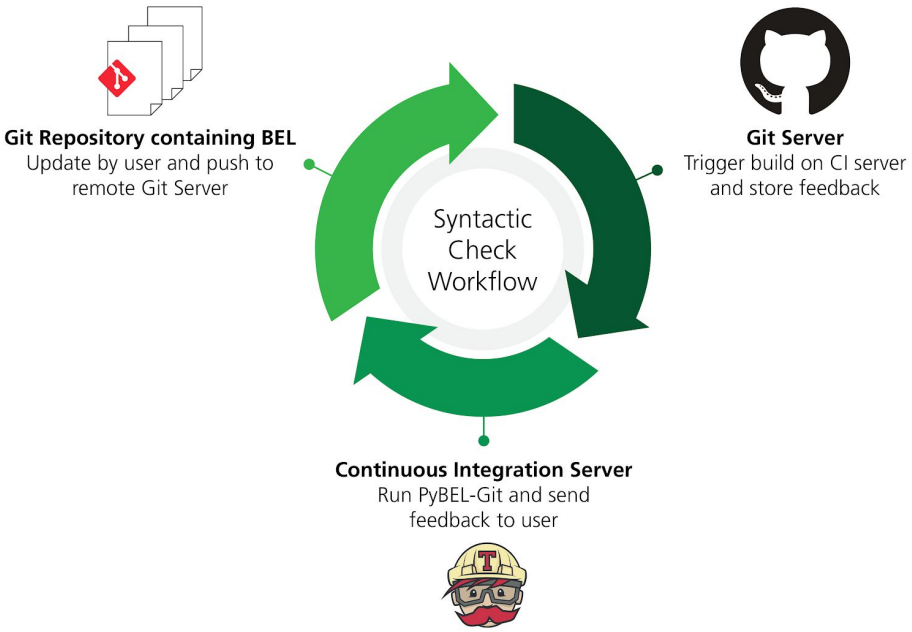


Figure 1: A workflow for syntactic quality assessment. This figure can be found on FigShare at https://doi.org/10.6084/m9.figshare.7643006.v1.

This workflow can be implemented in other web-based version control systems such as GitLab (https://gitlab.com) and Atlassian BitBucket (https://bitbucket.org) as well as directly integrated with continuous integration systems such as GitLab CI/CD (https://docs.gitlab.com/ee/ci), Travis-CI (https://travis-ci.com), and BitBucket Pipelines (https://bitbucket.org/product/features/pipelines) using the instructions provided at https://github.com/pybel/pybel-git with minimal configuration.

Semantic Quality Assurance

We selected ten signatures (and their corresponding BEL documents) from NeuroMMSig based on their druggability (number of proteins targeted by drugs that have been assessed in clinical trials), their novelty (less preference given to subgraphs corresponding to hypotheses that have repeatedly failed in the clinic; namely amyloid-beta aggregation), and their amenability to assay development (based on expert advice) as an example for the re-curation workflow outlined below. An enumeration and statistics can be found in **Table 1** and the signatures can be explored through BEL Commons (Hoyt *et al.*, 2018).

Label	Description	Before Re-curation		After Re-curation		After Enrichment	
		Nodes	Edges	Nodes	Edges	Nodes	Edges
Tau protein subgraph	The downstream effects of the post-translational modification, aggregation, and transport of the Tau protein	191	493	261	733	708	2054
DKK1 Subgraph GSK3 Subgraph	The interaction partners with GSK-3β and its targets of post-translational modification. The complementary DKK1 pathway is a specific signaling cascade upstream of GSK-3β	128	254	174	377	376	1165
Inflammatory Response	Processes related to inflammation in the context of Alzheimer's disease	182	373	341	743	2003	7607
Insulin Signal Transduction	The molecular relationships between insulin resistance and inflammation, motivated by epidemiological studies that suggested a correlation between AD and Type II diabetes (Karki *et al.*, 2017).	251	739	315	881	612	1973
Amyloidogenic Subgraph	The downstream effects of the amyloid precursor protein (APP), its protein modifiers, and its cleavage products	493	1223	652	1751	2090	7436
Non-amyloidogenic Subgraph	Chemicals and processes known to down-regulate the expression of the transcript corresponding to APP or the abundance of the APP protein	195	359	325	635	795	2238
Apoptosis and Cell Death	Processes relevant to AD that result in apoptosis including the Caspase subgraph, XIAP subgraph, and Complement system subgraph	104	143	170	229	1065	2401
Acetylcholine Subgraph	Pathways including biological entities and processes related to cholinergic neurons and acetylcholine transmission	106	197	148	337	423	1275
GABA Subgraph	Pathways including biological entities and process related to GABAergic neurons and GABA transmission	21	30	91	190	305	721
Reactive Oxygen Species Subgraph	The effects of reactive oxygen species, including the Myeloperoxidase subgraph, Hydrogen peroxide subgraph, Free radical formation subgraph, and Nitric oxide subgraph	104	173	126	224	1401	6277
Total		**1188**	**3529**	**1704**	**5391**	**5850**	**23811**

Table 1: Statistics for the number of BEL nodes and BEL statements in the ten knowledge graphs selected from the NeuroMMSig inventory before re-curation (using the version last updated on December 6th, 2016), after-recuration, and after enrichment. Later, we discuss these statistics in terms of INDRA statements - the discrepancies are due to the ontological reasoner applied in the conversion process from INDRA statements to BEL statements.

Because BEL was developed by the biomarker discovery company, Selventa, before the wide adoption of semantic resources like Identifiers.org, the Open Biomedical Ontology (OBO) Foundry, and the OLS, the language used a custom format for storing the names and identifiers of entities in major biomedical databases and ontologies such as the HUGO Genome Nomenclature Consortium (HGNC; Yates *et al.*, 2017) Chemical Entities of Biological Interest (ChEBI; Hastings *et al.*, 2013), the Gene Ontology (GO; Carbon *et al.,* 2017), Medical Subject Headings (MeSH; Rogers, 1963), the Disease

Ontology (DO; Schriml *et al.*, 2018), the Human Phenotype Ontology (HPO; Köhler *et al.*, 2018), the Cell Line Ontology (CLO; Sarntivijai *et al.*, 2014), the Experimental Factor Ontology (EFO; Malone *et al.*, 2010), and others. Additionally, Selventa provided several entity type-specific, manually curated terminologies for chemicals, protein families, protein complexes, and diseases for entities that had not yet been included in any of the other existing resources.

Because the Selventa terminologies are no longer maintained and the publicly available terminologies have far surpassed them in coverage, the first step in re-curation was to normalize entities to high-quality, publicly available terminologies. For example, chemicals were normalized to identifiers from ChEBI, ChEMBL (Gaulton *et al.*, 2017), and PubChem (Kim *et al.*, 2016) whenever possible; protein families and complexes were normalized to FamPlex (Bachman *et al.*, 2018); and diseases were normalized to DO and HPO. Further, because the BEL documents from AETIONOMY were all produced before 2015, the entities that were curated using their labels (instead of stable identifiers) needed to be updated. A short investigation showed that HGNC and GO were the least stable namespaces, but combined they had less than one hundred entities to be addressed. We therefore concluded that manual intervention was more appropriate than developing complicated systems for updating labels. While it is not intended to be the focus of this article, we have also begun to build a custom terminology (available at https://github.com/pharmacome/terminology) to supplement the publicly available ones for a small number (less than 1000) of terms that had not been included in other resources.

After ensuring both the correctness of BEL syntax and namespace usage, a remaining major aspect of re-curation is to address the issues arising from curation lacking inter-annotator agreement. BEL statements and their corresponding annotations (metadata) were generated by several independent curators and had not undergone quality control either by comparison with the results of independent curation of the same document by a second curator, or even minimally checked by a second curator. We applied the following simple guidelines:

1. *Second Curator:* check and label all relevant statements with a `SET Confidence` annotation using the Likert scale as described in **Table 2**.
2. *Third Curator (curation leader)*: after all relevant statements had been checked for correctness, check all statements with `SET Confidence = "High"` or `SET Confidence = "Medium"`. Change the confidence to `SET Confidence = "Very High"` on agreement. Otherwise, fix the statement.

Confidence	Rationale
None	If the evidence string is nonsense or contains no reasonable biological knowledge, delete it and the related statements entirely. It's okay to remove BEL statements that are not supported.
Low	If it's not clear what BEL should represent the biology, add `SET Confidence = "Low"` for later discussion.
Medium	If the statement is wrong, fix it and add the annotation `SET Confidence = "Medium"`.
High	If statement can be asserted from the given evidence, add the annotation `SET Confidence = "High"`.

Table 2. Confidence annotations using the Likert scale for re-curation

The existence of the confidence guideline can be checked with the PyBEL command line interface with the following command: `pybel compile --required-annotations "Confidence"`.

Proposed Approach for Updating and Rational Enrichment

Next, we developed and applied a procedure for enriching a given BEL document in order to cope with the mounting issues of out-of-dateness and incompleteness. Our approach identifies nodes with low information density and uses a large-scale corpus of biomedical literature that has been pre-processed by automated relation extraction methods to identify the most relevant literature, evidences, and ultimately relations. Notably, the previously described quality assurance (i.e., re-curation) workflows for checking and addressing the syntactic and semantic correctness of a given BEL document were necessary to decrease the noise input into the procedure. Following the re-curation of the ten NeuroMMSig subgraphs, we applied the following procedure for rational enrichment:

1. *Knowledge Graph Pre-processing*: nodes corresponding to the same gene (i.e., RNA, microRNA, Protein, and variants thereof) are collapsed, non-causal relationships (e.g., correlative, associative, ontological, etc.) are removed, and several entity types (i.e., abunances, reactions, pathologies, and biological processes) are removed.
2. *Application of Information Density Metric*: the remaining nodes are ranked by an information density function. We used the sum of the node in-degree and out-degree as this corresponds to the amount of causal information for a given gene in the knowledge graph. In this scenario, isolated nodes correspond to genes for which there is no causal information about its interactions with other proteins, and leaves (i.e., entities with only one edge) correspond to nodes that have very limited information.
3. *Automated Relation Extraction*: the top-ranked genes are used as a query to a knowledge graph generated by large-scale automated biological relation extraction. We used the Integrated Network and Dynamical Reasoning and Assembler (INDRA; Gyori *et al.*, 2018) and applied

several filters to find the most relevant and novel relations. First, the relations that were already curated and in the knowledge graph were excluded. Second, INDRA was used to calculate a confidence score (between 0.0 and 1.0) for each relation based on evidences from structured databases and the frequency of occurrence of similar statements. Those statements with a low confidence score (< 0.80) were removed to increase the precision and therefore reduce the curation overhead. While INDRA integrates relations extracted from multiple reading systems, a corpus of relations from a single machine reading system, such as EVEX, would serve the same purpose (Van Landeghem *et al*., 2012).

4. *Conversion to BEL*: different automated relation extraction systems present various information (e.g., entity offsets, events, triggers, etc.) in ways that are not amenable to curation. Because INDRA already normalizes this information for several systems to several varieties of the `indra.Statement` Python class, we developed a converter to BEL using PyBEL that can be used directly with the `indra.assemblers.PybelAssembler` Python class. Finally, this information is exported to an Excel sheet with several additional columns for tracking INDRA statement provenance, curator provenance, the correctness of BEL statements, the type of errors found, and the changes made to incorrect BEL statements. Examples and links to the full results can be found in the supplementary information.

For each round of rational enrichment, the procedure was applied to generate several curation sheets corresponding to the lowest information genes. Each row was checked with the following procedure:

1. Place an "x" in the *Checked* column.
2. If the BEL statement correctly corresponds to the *Evidence* column, place an "x" in the *Correct* column.
3. Else if the BEL statement can be improved (e.g., assignment of entity types, relation, etc.), correct it and place an "x" in the *Changed* column and annotate the error type in the *Error Type* column using a controlled vocabulary (see the supplementary data). Additional guidelines for categorizing error types can be found at https://github.com/pharmacome/curation/blob/master/indra-errors.rst.
4. Else if the BEL statement does not correspond to the *Evidence* column and can not be improved, then "x" should neither be placed in the *Correct* nor the *Changed* column.
5. If the *Evidence* column contains other BEL statements that were not extracted, duplicate the current row's provenance (reference, evidence, etc.) and add the additional BEL statements. Place an "x" in the *Changed* column but not the *Correct* column.

6. If there are other BEL statements that can be extracted, make a new line with all of the same provenance information (uuid, reference, evidence, etc.) and just place an "x" in the "Changed" column.

This procedure was applied iteratively: as the low information density nodes from the first round gained new relations, the knowledge graph was expanded and further low information density nodes were added. There are several improvements that could be made to the information density function and prioritization of the resulting extracted statements. For example, relations found by INDRA between low information density nodes and high information density nodes could be prioritized to maintain the scope and focus of a knowledge graph.

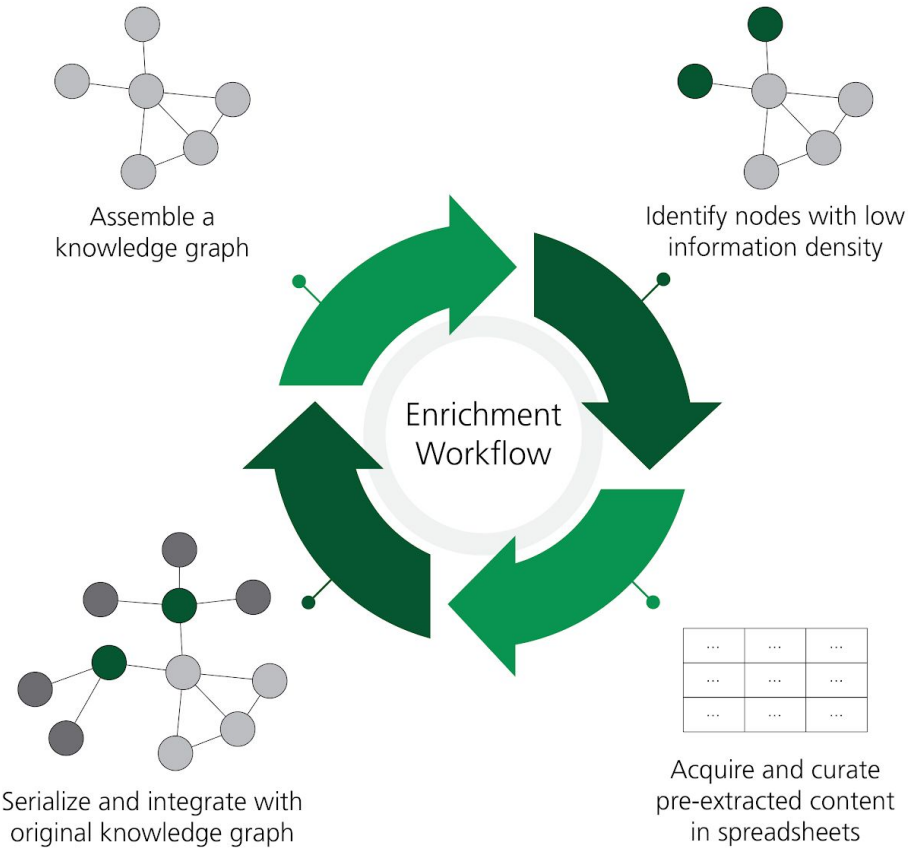


Figure 2. A workflow for the rational enrichment of knowledge graphs. This figure can be found on FigShare at https://doi.org/10.6084/m9.figshare.7642964.v1.

Results and Discussion

While applying the re-curation workflow outlined in **Figure 1**, we identified large sections of poor quality curation that had to be removed. Additionally, some evidences in the BEL document that were previously incompletely curated were completed. Re-curation also required the updating of namespaces from the 2015 versions to the most current and necessitated some additional revisions.

To evaluate the enrichment workflow outlined in **Figure 2**, we defined weekly curation rounds in which each of the five curators were tasked to curate the enrichment template generated by INDRA for the first 30 prioritized genes. Curators worked 10 hours per round for one month (4 weeks; one round per week) to curate BEL statements from a pool of 113 genes. A database of statements was generated by INDRA using the REACH (Valenzuela-Escárcega *et al.*, 2015; Valenzuela-Escárcega *et al.*, 2018), and Sparser (McDonald, 2000) readers to extract a total of 17096 statements containing these genes from all MEDLINE abstracts and PubMed Central full text articles available in August 2018. Of these, 2989 were manually evaluated. 917 statements (30.7%) were marked as correct by the curators, 1454 statements (48.6%) required manual corrections, and the remainder (20.7%) could not be corrected. The criteria for correctness was that *all* aspects of the statement, including the subject and object entities, relationship type, phosphorylation and other post-translational modifications, were extracted to the same extent as careful manual curation could. Ultimately, excluding the statements that could not be corrected, 79.3% of the automatically extracted, manually revised BEL statements were recovered. After curation, the recovered statements were converted into a BEL knowledge graph that contained 4228 nodes and 17002 edges complementary to the original ten subgraphs selected from NeuroMMSig. The discrepancies in the number of INDRA statements to BEL statements is due to the ontological reasoning process that occurs during conversion. For example, INDRA statements about protein complex formation are converted to bi-directional BEL statements, INDRA statements about post-translationally modified proteins induce edges to the reference protein, and INDRA statements about bound proteins create a variety of additional BEL nodes representing their constituents and membership edges connecting them.

There are two main aspects that are commonly used to formally evaluate a biocuration workflow: the time required to complete the task and quality of the curation compared with a gold standard. To evaluate whether the proposed approach for rational enrichment allows curating a larger amount of statements without compromising the quality, we calculated the average number of minutes required to curate one statement using our proposed workflow and compared it with previous estimates calculated conducting manual curation of BEL statements (Szostak *et al.*, 2015; Madan *et al.*, 2016) **(Figure 3a)**. While the

average curation effort was significantly lower than manual curation (2.19 minutes per BEL statement in our workflow vs. 3.2 minutes per BEL statement in manual curation), our calculations included the time used by the curators to annotate the various errors made by the reading system(s). Therefore, if the curation exercise would have exclusively focused on curating BEL statements, the average would have been even lower. Moreover, it is important to note that our proposed approach does not explicitly require the time nor expertise required for corpora generation because the reading systems (e.g., REACH and Sparser) and assembly systems (i.e., INDRA and PyBEL) are applied to all available literature.

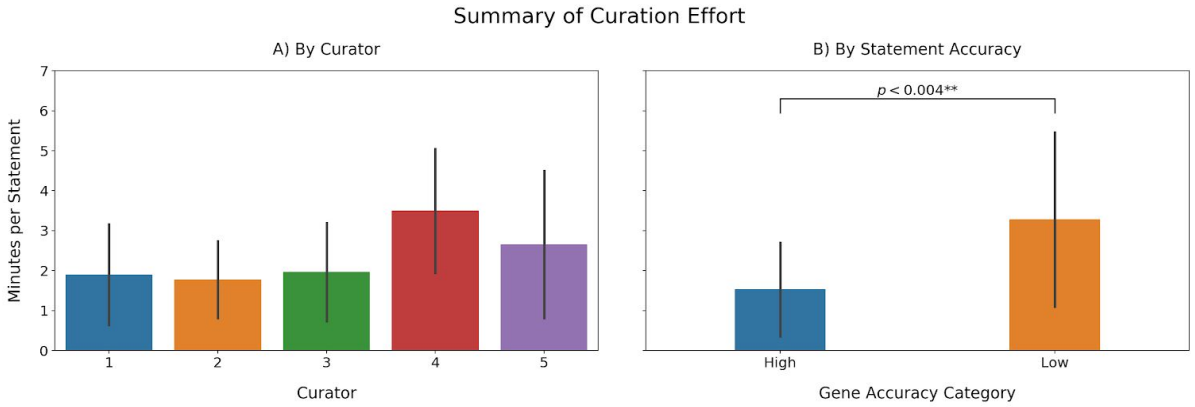


Figure 3. a) Recovered BEL statements per minute. Note that the time reported here includes the time invested in annotate the statement as well as INDRA errors. **b)** A comparison of the curation effort between genes for which INDRA had high accuracies (top 20) and genes presenting low accuracies (bottom 20).

Although the amount of time required to curate a certain amount of statements with the proposed approach is lower compared to standard manual curation, the curation effort is also highly variable depending on which gene was curated (**Figure 3a).** To investigate how the curation effort depends on the accuracy of the reader extracting BEL statements, we compared the average curation effort between genes whose statements were accurately and poorly extracted **(Figure 3b)**. We observed that the curation effort required to extract statements in genes whose statements were highly accurate (top 20) was significantly less ($p < 0.004$; Student's T) than the effort required to curate low accuracy (bottom 20) genes, which effectively took as long as manual curation. We conclude that the high variability associated with the average curation times per curator can be explained by the extra invested time in the genes presenting low recall.

The second aspect we evaluated was the performance in terms of quality. To investigate the direct quality of the BEL statements coming from INDRA, we analyzed the distributions of correct statements before curation observed in each gene (accuracy investigation) **(Figure 4a)**. Most of the genes presented

accuracies close to the mean accuracy (35.75%) with only a few outliers whose limited number of extracted statements lead to their respective high or low accuracies **(see Supplementary Figure 1)**. Furthermore, in accordance with previous research assessing the quality of automatic and manual relation extraction (Rinaldi *et al.,* 2016), the accuracies we observed again indicated that BEL statements must be manually curated in order to generate high quality networks. After curation, the distribution of statements that were correct plus statements that were fixed during curation (i.e., excluding statements that were incorrect and could not be fixed) shifted completely to long-tailed distribution with an average of 74.63% BEL statements successfully extracted **(Figure 4b)**. The remaining statements (approximately 25%) could either not be coded in BEL nor contained any relevant information about the particular gene.

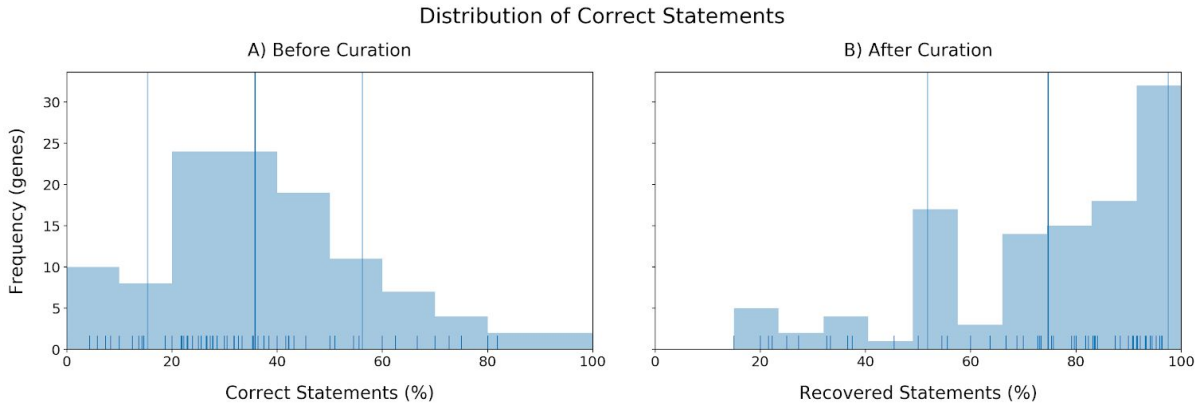


Figure 4. a) The distribution of the accuracies in triple identification by INDRA for each gene. X-axis: Correct statements (%). Y-axis: Number of genes (frequency). **b)** Distribution of recovered statements after curation (mean: 74.63%).

While curating the BEL statements, we also annotated the errors made throughout the process of reading, assembly by INDRA, and conversion to BEL by PyBEL in order to identify common mistakes and to assist in the improvement of these three systems. The results showed that the most common error is caused by the name-entity recognition system that identifies the entities participating in the relation **(Figure 5).** Other common errors arose from the improper assignment of the subject and object entities, from evidences that did not actually include relations between the subject and object entities, and statements that were semantically incorrect due to a negation word (e.g., not, no, none, neither, etc.).

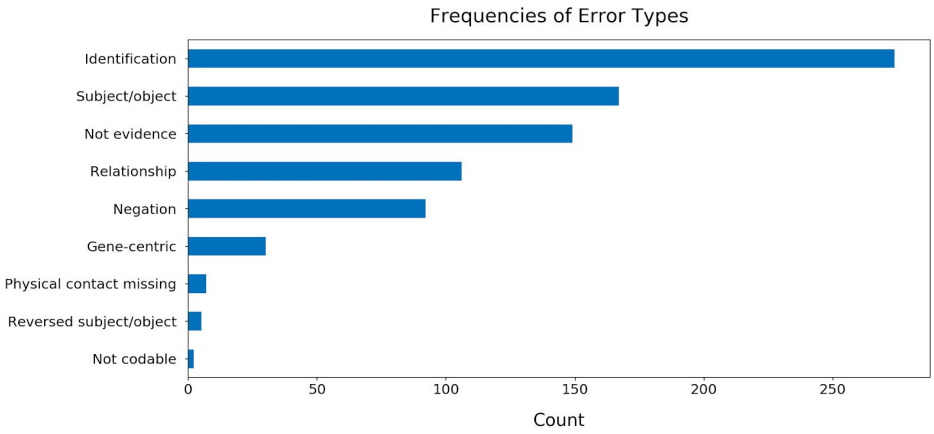


Figure 5. The frequencies of common errors found while curating BEL statements generated from 113 genes. Further details about each error type and the annotation process are available in the guidelines available at https://github.com/pharmacome/curation/blob/master/indra-errors.rst.

The five curators were tasked with tagging interesting examples of the common mistakes that could be used to inform the development of the reading systems (REACH, Sparser, etc.) and the assembly systems (INDRA and PyBEL). Because the authors of this manuscript maintain the INDRA and PyBEL packages, identifying the causes of errors in assembly was relatively straightforward. For example, BEL statements containing biological processes were consistently output using invalid BEL syntax, including the *activity()* function, which is reserved for proteins and other physical entities. We addressed this by updating the previously mentioned *indra.assemblers.PybelAssembler* class. Another error type that was not addressed until after the evaluation was completed was the determination of the role of direct physical interaction in causal relations. INDRA makes use of linguistic cues from the text mining systems along with information from protein-protein interaction databases to determination if a relation involves a physical interaction between proteins, but this information was not incorporated into the *indra.assemblers.PybelAssembler* class. Instead, by default all relations were output using BEL statements implying physical contact: "directly increases" (i.e. increases via contact) and directly decreases (i.e., decreases via contact). This issue has since been fixed. In general, the direct/indirect distinction is difficult to detect automatically in natural language, though it is very important in the generation of mechanistic and mathematical models arising from biological knowledge.

In **Table 3**, we present a small sampling of the errors and corresponding suggestions for improvement in the reading systems. We present a much more thorough enumeration of the errors found in statements for the 113 curated genes in the supplementary information. Besides generating new content quickly, this curation procedure includes information to allow for the evaluation of the automated relation extraction systems and for the proposition of improvements. For example, new groundings can be proposed for

entities that were often mismatched. A prominent example was the misidentification of tau (a human protein) and taurine (an amino acid).

Additionally, new rules could be suggested for rule-based systems to avoid issues with the mis-identification of the order of the subject and object as in the example of *"Bak expression was also induced in cells overexpressing the stress-induced transcription factor GADD153, but Bak expression was inhibited in cells expressing an antisense GADD153 construct"* (Lovat *et al.*, 2003) whose use of the passive voice may have caused REACH to interpret the statement as "*Bak increased GADD153*." Ultimately, we believe we can use these examples to provide useful feedback to the developers of the reading systems and improve future extraction.

Gene	Evidence	Issue	Suggestion
MRC1	In conclusion, these results suggest that BCR and ABL kinase abrogates MMR activity to inhibit apoptosis and induce mutator phenotype. (Stoklosa *et al.*, 2008)	MRC1, also known as MMR, was confused with Mismatch repair (MMR)	Machine learning methods generating contextual word embeddings could be used to improve the named entity recognition component such as NeuralCoref (https://github.com/huggingface/neuralcoref)
TIMP1	In our work, the restoration of cholesterol efflux capacities from EPA enriched HMDM treated with both the adenylate cyclase activator forskolin and the phosphodiesterase inhibitor IBMX strongly suggests that EPA decreased the ABCA1 mediated cholesterol efflux from HMDM through a PKA dependent pathway. (Fournier *et al.,* 2016)	TIMP1, also known as EPA, was confused with eicosapentaenoic acid (EPA)	Improve the named entity recognition (disambiguation) process, for example, by updating synonym dictionaries in rule-based systems.
TRPV1	Moreover, recently TRPV1 has been demonstrated to be either inhibited or activated by PIP 2. (Morelli *et al.*, 2014)	Only the inhibition relationship was extracted	Rule-based relation extraction systems could be appended with new rules to handle sentences with multiple objects. This and similar examples could be included in the training data for machine learning-based relation extraction.
NUMB	This interaction is mediated by the NPXY motif of LNX1 and leads to ubiquitination of Numb by the RING domain of LNX1, thereby targeting Numb to proteasomal degradation. (Young *et al.,* 2005)	The complex sentence structure of "ubiquitination" and "targeting" event were not resolved properly, and the ubiquitination was omitted.	Rule-based systems like REACH that explicitly handle ubiquitination events could be appended with new rules.
USF2	Taken together, the results shown in Figs. 5A, B and C suggest that USF2 stimulates the transcriptional activity of NFκB by enhancing the degradation of IκBα. (Wand *et al.*, 2009)	Relation should be treated as an indirect, rather than direct, increase	Update the INDRA *PybelAssembler* to make use of information about whether a relation is mediated through physical contact.

Table 3: Examples of errors that resulted in suggestions for improvements for the underlying relation extraction systems.

After applying the re-curation workflow to our selection of knowledge graphs in the NeuroMMSig inventory, we increased the number of nodes from 1188 to 1704 (~1.5x) and edges from 3529 to 5391 (~1.5x). After applying the enrichment workflow, the number of nodes increased to 5850 (~5x) and edges to 23811 (~7x). A more granular summary can be found in **Table 1**. With a 5x increase in nodes, we would expect to see a 10x increase in edges if the new nodes were completely disconnected from the pre-existing nodes in the knowledge graph, which shows that we have been able to maintain the specificity of the knowledge graphs to a reasonable degree. In total, our curators spent 80 hours on the enrichment step to generate 17,002 new BEL statements with an average rate of 3.54 edges per minute. The resulting enriched knowledge graph can be used in reproductions of previous analyses leveraging the NeuroMMSig inventory to assess their robustness, deliver new insights, and improve future analyses when the results are incorporated into a future release of the NeuroMMSig mechanism enrichment server. Additionally, the statements comprise a large training set for future machine learning approaches for text mining.

Conclusions

We have proposed and applied a generalizable workflow for enriching and updating existing biological knowledge graphs with a focus on the reduction of curation time both in literature triage and in extraction. While its realization involved spreadsheets rather than a *bona fide* curation interface, we believe that it could be adopted by both BEL-specific curation interfaces (e.g., BELIEF, BioDati Studio[1]) and more general biological relation curation interfaces (e.g., NOCTUA[2], Factoid[3], WikiPathways (Slenter *et al.*, 2017)). Furthermore, INDRA is flexible enough to generate curation sheets for curators familiar with formats other than BEL, such as BioPAX or SBML.

This workflow is by no means the ultimate solution for finding relevant content. Using pre-extracted statements as a stand-in for relevance allows a given knowledge graph to be expanded, but it requires several rounds to find the limits of a given pathway or graph, during which the scope of the curation could be lost. We plan to investigate other methods for identifying relevant content by combining topic modeling with mind maps to not only identify content at the entity level, but on a higher abstraction that allows for capturing of entire areas of biology. These methods could compensate for the simplications that

[1] https://studio.demo.biodati.com
[2] http://noctua.berkeleybop.org
[3] https://github.com/PathwayCommons/factoid

we made to the curation task, such as removing relations containing chemicals, biological processes, and phenotypes. Additionally, they could enable earlier-stage curation that is more focused on achieving reasonable coverage of the available knowledge rather than high granularity enrichment.

Ultimately, as automated relation extraction technologies improve, they will be used to more significantly supplement manual curation efforts. We expect to see many upcoming workflows leveraging these exciting prospects.

Declarations

Acknowledgements

We would like to thank Stephan Gebel for his organizational support and Alina Enns and Keerthika Lohanadan for their help in the curation tasks.

Funding

This work was supported by the Fraunhofer Society under the MAVO project, the Human Brain Pharmacome (https://pharmacome.scai.fraunhofer.de). D.D.F 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.

Authors' Contributions

C.T.H. and D.D.F. conceived and designed the study and authored this manuscript. C.T.H., D.D.F., R.A., L.X., S.S., E.W., and K.K. performed curation. J.B., B.G., and P.G. provided data. M.H.A. supervised the project.

Availability of Data and Materials

The pybel-git Python package that was used to assess syntactic quality is openly available at https://github.com/pybel/pybel-git. All other code and analysis is openly available at https://github.com/bel-enrichment.

Competing Interests

The authors declare that they have no competing interests.

References

1. Bachman, J. A., Gyori, B. M., & Sorger, P. K. (2018). FamPlex: A resource for entity recognition and relationship resolution of human protein families and complexes in biomedical text mining. *BMC Bioinformatics*, *19*(1), 1–14. https://doi.org/10.1186/s12859-018-2211-5
2. Belleau, F., *et al.* (2008). Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. *Journal of Biomedical Informatics*, *41*(5), 706–716. https://doi.org/10.1016/j.jbi.2008.03.004
3. Carbon, S., *et al.* (2017). Expansion of the gene ontology knowledgebase and resources: The gene ontology consortium. *Nucleic Acids Research*, *45*(D1), D331–D338. https://doi.org/10.1093/nar/gkw1108
4. Caspi, R., *et al.* (2016). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Research*, *44*(D1), D471–D480. https://doi.org/10.1093/nar/gkv1164
5. Cerami, E. G., *et al.* (2011). Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Research*, *39*(SUPPL. 1), 685–690. https://doi.org/10.1093/nar/gkq1039
6. Cote, R., Jones, P., Apweiler, R., & Hermjakob, H. (2006). The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries. *BMC Bioinformatics*, *7*, 1–7. https://doi.org/10.1186/1471-2105-7-97
7. Demir, E., *et al.* (2010). The BioPAX community standard for pathway data sharing. *Nature Biotechnology*, *28*(12), 1308–1308. https://doi.org/10.1038/nbt1210-1308c
8. Domingo-Fernández, D., Hoyt, C. T., Bobis Alvarez, C., Marin-Llao, J., & Hofmann-Apitius, M. (2018). ComPath: An ecosystem for exploring, analyzing, and curating pathway databases. *npj Systems Biology and Applications, 5(1), 3. https://doi.org/10.1038/s41540-018-0078-8*
9. Domingo-Fernández, D., Mubeen, S., Marin-Llao, J., Hoyt, C., & Hofmann-Apitius, M. (2019). PathMe: Merging and exploring mechanistic pathway knowledge. *bioRxiv*. Retrieved from http://biorxiv.org/content/early/2018/10/24/451625
10. Domingo-Fernández, *et al.* (2017). Multimodal mechanistic signatures for neurodegenerative diseases (NeuroMMSig): A web server for mechanism enrichment. *Bioinformatics*, *33*(22), 3679–3681. https://doi.org/10.1093/bioinformatics/btx399
11. Emon, M. A. E. K., Kodamullil, A. T., Karki, R., Younesi, E., & Hofmann-Apitius, M. (2017). Using Drugs as Molecular Probes: A Computational Chemical Biology Approach in

Neurodegenerative Diseases. *Journal of Alzheimer's Disease*, *56*(2), 677–686. https://doi.org/10.3233/JAD-160222

12. Fournier, N., *et al.* (2016). Eicosapentaenoic acid membrane incorporation impairs ABCA1-dependent cholesterol efflux via a protein kinase A signaling pathway in primary human macrophages. *Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids*, *1861*(4), 331-341.
13. Gaulton, A., *et al.* (2017). The ChEMBL database in 2017. *Nucleic Acids Research*, *45*(D1), D945–D954. https://doi.org/10.1093/nar/gkw1074
14. Glont, M., *et al.* (2018). BioModels: Expanding horizons to include more modelling approaches and formats. *Nucleic Acids Research*, *46*(D1), D1248–D1253. https://doi.org/10.1093/nar/gkx1023
15. Gonçalves, R. S., *et al.* (2017). The CEDAR workbench: An ontology-assisted environment for authoring metadata that describe scientific experiments. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10588 LNCS*, 103–110. https://doi.org/10.1007/978-3-319-68204-4_10
16. Guryanova, S., & Guryanova, A. (2017). sbv IMPROVER: Modern Approach to Systems Biology. *Methods in Molecular Biology (Clifton, N.J.)*, *1613*, 21–29. https://doi.org/10.1007/978-1-4939-7027-8_2
17. Gyori, B. M., *et al*. (2017). From word models to executable models of signaling networks using automated assembly. *Molecular Systems Biology*, *13*(11), 954. https://doi.org/10.15252/msb.20177651
18. Hastings, J., *et al.* (2013). The ChEBI reference database and ontology for biologically relevant chemistry: Enhancements for 2013. *Nucleic Acids Research*, *41*(D1), 456–463. https://doi.org/10.1093/nar/gks1146
19. Hofmann-Apitius, M., *et al.* (2015). Bioinformatics mining and modeling methods for the identification of disease mechanisms in neurodegenerative disorders. *International journal of molecular sciences*, *16*(12), 29179-29206. https://doi.org/10.3390/ijms161226148
20. Hoyt, C.T. (2018). cthoyt/pybel-git v0.0.1 (Version v0.0.1). Zenodo. http://doi.org/10.5281/zenodo.1491432
21. Hoyt, C. T., Domingo-Fernández, D., & Hofmann-Apitius, M. (2018). 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, 1 January 2018, bay126, https://doi.org/10.1093/database/bay126

22. Hoyt, C. T., Domingo-Fernández, D., Balzer, N., Güldenpfennig, A., & Hofmann-Apitius, M. (2018). A systematic approach for identifying shared mechanisms in epilepsy and its comorbidities. *Database : The Journal of Biological Databases and Curation*, *2018*(June), 269860. https://doi.org/10.1093/database/bay050
23. Hoyt, C. T., Konotopez, A., & Ebeling, C. (2018). PyBEL: a computational framework for Biological Expression Language. *Bioinformatics (Oxford, England)*, *34*(4), 703–704. https://doi.org/10.1093/bioinformatics/btx660
24. Hucka, M., *et al.* (2003). The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, *19*(4), 524–531. https://doi.org/10.1093/bioinformatics/btg015
25. Irin, A.K., Tom Kodamullil, A., Gündel, M., & Hofmann-Apitius, M. (2015). Computational Modelling Approaches on Epigenetic Factors in Neurodegenerative and Autoimmune Diseases and Their Mechanistic Analysis. *Journal of Immunology Research*, *2015*, 1–10. https://doi.org/10.1155/2015/737168
26. Juty, N., Le Novère, N., & Laibe, C. (2012). Identifiers.org and MIRIAM Registry: Community resources to provide persistent identification. *Nucleic Acids Research*, *40*(D1), 580–586. https://doi.org/10.1093/nar/gkr1097
27. Kamburov, A., *et al*. (2013). The ConsensusPathDB interaction database: 2013 Update. *Nucleic Acids Research*, *41*(D1), 793–800. https://doi.org/10.1093/nar/gks1055
28. Kandasamy, K., *et al*. (2010). NetPath: a public resource of curated signal transduction pathways. *Genome Biology*, *11*(1), R3. https://doi.org/10.1186/gb-2010-11-1-r3
29. Karki, R., Tom Kodamullil, A., & Hofmann-Apitius, M. (2017). Comorbidity Analysis between Alzheimer's Disease and Type 2 Diabetes Mellitus (T2DM) Based on Shared Pathways and the Role of T2DM Drugs. *Journal of Alzheimer's Disease*, *60*(2), 721-731.
30. Kim, S., *et al.* (2016). PubChem substance and compound databases. *Nucleic Acids Research*, *44*(D1), D1202–D1213. https://doi.org/10.1093/nar/gkv951
31. Kodamullil, A. T., Younesi, E., Naz, M., Bagewadi, S., & Hofmann-Apitius, M. (2015). Computable cause-and-effect models of healthy and Alzheimer's disease states and their mechanistic differential analysis. *Alzheimer's and Dementia*, *11*(11), 1329–1339. https://doi.org/10.1016/j.jalz.2015.02.006
32. Köhler, S., Carmody, L., Vasilevsky, N., Jacobsen, J. O. B., Danis, D., Gourdine, J.-P., … Robinson, P. N. (2018). Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Research*, 1–10. https://doi.org/10.1093/nar/gky1105

33. Laibe, C., & Le Novère, N. (2007). MIRIAM Resources: tools to generate and resolve robust cross-references in Systems Biology. *BMC Systems Biology*, *1*, 58. https://doi.org/10.1186/1752-0509-1-58
34. Lovat, P. E., *et al.* (2003). Bak: a downstream mediator of fenretinide-induced apoptosis of SH-SY5Y neuroblastoma cells. *Cancer Research*, *63*(21), 7310–3.
35. Madan, S., *et al.* (2016). The BEL information extraction workflow (BELIEF): evaluation in the BioCreative V BEL and IAT track. *Database : The Journal of Biological Databases and Curation*, *2016*(September 2017), 1–17. https://doi.org/10.1093/database/baw136
36. Malone, J., *et al*. (2010). Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics*, *26*(8), 1112–1118. https://doi.org/10.1093/bioinformatics/btq099
37. McDonald, D. D. (2000). Issues in the Representation of Real Texts: The Design of Krisp. *Natural Language Processing and Knowledge Representation*, 77–110.
38. Meldal, B. H. M., *et al*. (2015). The complex portal - An encyclopaedia of macromolecular complexes. *Nucleic Acids Research*, *43*(D1), D479–D484. https://doi.org/10.1093/nar/gku975
39. Mihindukulasooriya, N., Hassanzadeh, O., Dash, S., & Gliozzo, A. (2017). Towards comprehensive noise detection in automatically-created knowledge graphs. *CEUR Workshop Proceedings*, *1963*, 1–4.
40. Morelli, M. B., *et al*. (2014). Cross-talk between alpha 1D-adrenoceptors and transient receptor potential vanilloid type 1 triggers prostate cancer cell proliferation. *BMC cancer*, *14*(1), 921.
41. Naz, M., Kodamullil, A. T., & Hofmann-Apitius, M. (2016). Reasoning over genetic variance information in cause-and-effect models of neurodegenerative diseases. *Briefings in Bioinformatics*, *17*(3), 505–16. https://doi.org/10.1093/bib/bbv063
42. Nickel, M., *et al.* (2016). A Review of Relational Machine Learning for Knowledge Graphs. Proceedings of the IEEE, 104(1), 11–33. https://doi.org/10.1109/jproc.2015.2483592
43. Perfetto, L., *et al*. (2016). SIGNOR: A database of causal relationships between biological entities. *Nucleic Acids Research*, *44*(D1), D548–D554. https://doi.org/10.1093/nar/gkv1048
44. Pilalis, E., *et al*. (2015). KENeV: A web-application for the automated reconstruction and visualization of the enriched metabolic and signaling super-pathways deriving from genomic experiments. *Computational and Structural Biotechnology Journal*, *13*, 248–255. https://doi.org/10.1016/j.csbj.2015.03.009
45. Pon, A., *et al*. (2015). Pathways with PathWhiz. *Nucleic Acids Research*, *43*(W1), W552–W559. https://doi.org/10.1093/nar/gkv399

46. Pujara, J., Augustine, E., & Getoor, L. (2017). Sparsity and Noise: Where Knowledge Graph Embeddings Fall Short. *Conference on Empirical Methods in Natural Language Processing*, 1752–1757.
47. Rausanu, S., *et al.* (2015). Computational models for inferring biochemical networks. *Neural Computing and Applications*, *26*(2), 299–311. https://doi.org/10.1007/s00521-014-1617-x
48. Rinaldi, F., *et al.* (2016). BioCreative V track 4: a shared task for the extraction of causal network information using the Biological Expression Language. *Database*, *2016*.
49. Rodriguez-Esteban, R. (2015). Biocuration with insufficient resources and fixed timelines. *Database*, *2015*(1), 1–9. https://doi.org/10.1093/database/bav116
50. Rogers, F. B. (1963). Medical subject headings. *Bulletin of the Medical Library Association*, *51*, 114–6.
51. Saqi, M., *et al.* (2018). Navigating the disease landscape: knowledge representations for contextualizing molecular signatures. *Briefings in bioinformatics*, bby025. https://doi.org/10.1093/bib/bby025
52. Sarntivijai, S., *et al*. (2014). CLO: The cell line ontology. *Journal of Biomedical Semantics*, *5*(1), 1–10. https://doi.org/10.1186/2041-1480-5-37
53. Schriml, L. M., *et al.* (2018). Human Disease Ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Research*, 1–8. https://doi.org/10.1093/nar/gky1032
54. Slater, T. (2014). Recent advances in modeling languages for pathway maps and computable biological networks. *Drug Discovery Today*, *19*(2), 193–198. https://doi.org/10.1016/j.drudis.2013.12.011
55. Slenter, D. N., Kutmon, M., Hanspers, K., *et al.* (2017). WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic acids research*, *46*(D1), D661-D667.
56. Stobbe, M. D., Houten, S. M., Jansen, G. A., van Kampen, A. H., & Moerland, P. D. (2011). Critical assessment of human metabolic pathway databases: a stepping stone for future integration. *BMC systems biology*, *5*(1), 165.
57. Stoklosa, T., *et al.* (2008). BCR/ABL inhibits mismatch repair to protect from apoptosis and induce point mutations. *Cancer Research*, *68*(8), 2576-2580.
58. Szostak, J.,*et al.* (2015). Construction of biological networks from unstructured information based on a semi-automated curation workflow. *Database*, *2015*.

59. Tripathi, S., *et al.* (2015). The gastrin and cholecystokinin receptors mediated signaling network: A scaffold for data analysis and new hypotheses on regulatory mechanisms. *BMC Systems Biology*, *9*(1), 1–15. https://doi.org/10.1186/s12918-015-0181-z
60. Valenzuela-Escárcega, M. A., Hahn-Powell, G., Hicks, T., & Surdeanu, M. (2015). A Domain-independent Rule-based Framework for Event Extraction. Proceedings of ACL-IJCNLP 2015 System Demonstrations, 127–132.
61. Valenzuela-Escárcega, M. A., *et al*. (2018). Large-scale automated machine reading discovers new cancer-driving mechanisms. *Database : The Journal of Biological Databases and Curation*, *2018*, 1–14. https://doi.org/10.1093/database/bay098
62. Van Landeghem, S., *et al.* (2012). Exploring biomolecular literature with EVEX: Connecting genes through events, homology, and indirect associations. *Advances in Bioinformatics*, *2012*. https://doi.org/10.1155/2012/582765
63. Wadi, L., Meyer, M., Weiser, J., Stein, L. D., & Reimand, J. (2016). Impact of outdated gene annotations on pathway enrichment analysis. *Nature Methods*, *13*(9), 705–706. https://doi.org/10.1038/nmeth.3963
64. Wang, L., *et al.* (2009). HINT1 inhibits β-catenin/TCF4, USF2 and NFκB activity in human hepatoma cells. *International journal of cancer*, *124*(7), 1526-1534.
65. Williams, A. J., *et al.* (2012). Open PHACTS: semantic interoperability for drug discovery. *Drug Discovery Today*, *17*(21–22), 1188–1198. https://doi.org/10.1016/j.drudis.2012.05.016
66. Wishart, D. S., *et al.* (2018). HMDB 4.0: The human metabolome database for 2018. *Nucleic Acids Research*, *46*(D1), D608–D617. https://doi.org/10.1093/nar/gkx1089
67. Yates, B., *et al*. (2017). Genenames.org: The HGNC and VGNC resources in 2017. *Nucleic Acids Research*, *45*(D1), D619–D625. https://doi.org/10.1093/nar/gkw1033
68. Young, Paul, *et al.* (2005). LNX1 is a perisynaptic Schwann cell specific E3 ubiquitin ligase that interacts with ErbB2. *Molecular and Cellular Neuroscience* 30.2, 238-248.
69. Yugi, K., Kubota, H., Hatano, A., & Kuroda, S. (2016). Trans-Omics: How To Reconstruct Biochemical Networks Across Multiple 'Omic' Layers. *Trends in Biotechnology*, *34*(4), 276–290. https://doi.org/10.1016/j.tibtech.2015.12.013

Conclusions

5 Integration of structured biological data sources using Biological Expression Language

Introduction

Integration of Structured Biological Data Sources using Biological Expression Language

Charles Tapley Hoyt[1,2,*], Daniel Domingo-Fernández[1,2], Sarah Mubeen[1,2], Josep Marín-Llaó[1], Andrej Konotopez[1], Christian Ebeling[1], Colin Birkenbihl[1,2], Özlem Muslu[1,2], Bradley English[1,2], Simon Müller[1,2], Mauricio Pio de Lacerda[2], Mehdi Ali[3,4], Scott Colby[5], Dénes Türei[6,7,8], Nicolàs Palacio-Escat[7,8], and Martin Hofmann-Apitius[1,2]

1. Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Sankt Augustin 53754, Germany
2. Bonn-Aachen International Center for IT, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn 53115, Germany
3. Department of Enterprise Information Systems, Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS), Sankt Augustin 53754, Germany
4. Department of Computer Science, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn 53115, Germany
5. Department of Chemistry, Stanford University, Stanford, CA 94305, United States of America
6. European Molecular Biology Laboratory (EMBL), Structural and Computational Biology Unit, Meyerhofstrasse 1, D-69117, Heidelberg, Germany
7. RWTH Aachen University, Faculty of Medicine, Joint Research Centre for Computational Biomedicine, MTZ Pauwelsstraße 19, D-52074, Aachen, Germany
8. Heidelberg University, Faculty of Medicine and Heidelberg University Hospital, Institute of Computational Biomedicine, Bioquant Im Neuenheimer Feld 267, 69120 Heidelberg, Germany

***Corresponding Author**: Hoyt, C. T., Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Sankt Augustin 53754, Germany. Telephone details; +49 2241 14-2268. Email: charles.hoyt@scai.fraunhofer.de

Keywords: Data Integration, Semantic Web, Biological Expression Language, Knowledge graphs

Tweet

A suite of independent Python packages for downloading, parsing, warehousing, and converting multi-modal and multi-scale biological databases to Biological Expression Language

Abstract

Background: The integration of heterogeneous, multiscale, and multimodal knowledge and data has become a common prerequisite for joint analysis to unravel the mechanisms and aetiologies of complex diseases. Because of its unique ability to capture this variety, Biological Expression Language (BEL) is well suited to be further used as a platform for semantic integration and harmonization in networks and systems biology.

Results: We have developed numerous independent packages capable of downloading, structuring, and serializing various biological data sources to BEL. Each Bio2BEL package is implemented in the Python programming language and distributed through GitHub (https://github.com/bio2bel) and PyPI.

Conclusions: The philosophy of Bio2BEL encourages reproducibility, accessibility, and democratization of biological databases. We present several applications of Bio2BEL packages including their ability to support the curation of pathway mappings, integration of pathway databases, and machine learning applications.

1. Background

The integration of heterogeneous, multi-scale, and multi-modal biomedical data has become a cornerstone of modern computational investigation of the mechanisms and aetiologies underlying complex diseases (Iyappan *et al.*, 2014; van Dam *et al.* 2014; Wanichthanarak *et al.*, 2015; Himmelstein et al., 2017; Fan *et al.*, 2019). An overarching strategy was proposed by Davidson *et al.* more than two decades ago that outlined the transformation of data into a common model, semantic alignment of related objects, integration of schemata, and federation of data (Davidson *et al.*, 1995). However, integration remains a challenging task that requires the identification and deep understanding of biological data sources and their respective formats, conversion, harmonization, and unification.

Initial interest in the semantic web and linked open data along with the adoption of RDF (Resource Description Framework[1]) in the biomedical community led to the Bio2RDF project, in which pipelines for converting and serializing several biological data sources to RDF were developed (Belleau *et al.*, 2008). Several updates have been issued since its deployment such as the inclusion of chemical information systems (Chen *et al.*, 2010). Further, it has also influenced in and has been adopted by subsequent projects such as Open PHACTS (Williams *et al.*, 2012). While RDF is highly expressive and each of these projects have developed and enforced well-defined schemata, the format is often not well-suited for downstream analyses and must first be queried with languages like SPARQL (SPARQL Query Language for RDF[2]) and subsequently be transformed into appropriate formats with general-purpose programming languages. Alternatives to RDF/SPARQL such as property graphs (e.g., Neo4j[3], OrientDB[4]) are comparable (Alocci *et al.*, 2015) but also necessitate similar post-processing.

Conversely, there have been several biologically meaningful integration efforts (e.g., STRING; Warde-Farley, *et al.* 2010, GeneMANIA; Szklarczyk *et al.*, 2015, GeneCards; Stelzer *et al.*, 2016). However, most suffer from a lack of defined schemata or standardized data format that impede biological database interoperability. As interoperability itself is a multifaceted concept, we would like to highlight three of its facets: first, data sources should refer to named entities using high-quality, publicly accessible terminologies as prescribed by the Minimal Information Requested in the Annotation of Biochemical Models standard (Laibe and Le Novère, 2007). Second, data sources should additionally denote the ontological classes of named entities (e.g., gene, transcript, protein, pathway, disease) along with their reference using controlled vocabularies such as the Systems Biology Ontology (Courtot *et al.*, 2011). Some identifiers, such as those for genes, are often used to refer not only to the physical region of DNA within the genome, but also the corresponding RNA transcript(s) or protein product(s). Unfortunately, many biological databases do not explicitly distinguish between these entity classes. For example, the STRING database lists gene-centric homology relationships, transcript-centric co-expression relationships, and protein-centric protein-protein interactions using gene-centric nomenclature. While it may be possible to identify the classes of entities based on their incident relationships, doing so requires specific knowledge of the database including the semantics of its relationships. Third, resources should, at a minimum, map their relationships to

[1] https://www.w3.org/RDF
[2] https://www.w3.org/TR/rdf-sparql-query
[3] https://neo4j.com
[4] https://orientdb.com

controlled vocabularies such as the Relation Ontology[5], or further use standardized data formats with defined semantics (e.g., PSI MI-TAB[6]) to minimize both the interpretation and implementation effort when combining them with other resources.

OmniPath (Türei *et al*., 2016) began to address these facets when it combined several signaling pathway and transcriptional regulation databases. It achieved interoperability between several databases by normalizing the identifiers and relationships between entities from several databases describing the same phenomena (e.g., microRNA-target interactions, protein-protein interactions, etc.) and creating a unified network. However, because it did not use a standard format or schema as mentioned in the third facet for interoperability, OmniPath itself cannot readily be directly integrated with other biological data sources. Pathway Commons (Cerami *et al.,* 2011) addressed this concern when combining several molecular pathway and interaction databases by translating the source databases into the BioPAX standard (Demir *et al.*, 2010) using automated pipelines. However, it suffers from low granularity and low recovery of information from some of its primary biological data sources which may be due to prioritization of software development time, data usage restrictions, or shortcomings in the BioPAX standard. While BioPAX is well-suited for representing biological reactions and transformations, it is limited in its ability to represent correlative and associative relationships across multi-scale biology (e.g., at the levels of processes, phenotypes, and clinical observations).

As an alternative, we propose the use of Biological Expression Language (BEL; Slater, 2014) as an integration schema in order to overcome the limits faced by previous efforts and to simultaneously address all three facets of interoperability. BEL has begun to prove itself as a robust format in the curation and integration of previously isolated biological data sources of high granular information on genetic variation (Naz *et al.*, 2016), epigenetics (Irin *et al.*, 2015), chemogenomics (Emon *et al.*, 2017), and clinical biomarkers (Iyappan *et al.*, 2017). Its syntax and semantics are also appropriate for representing, for example, disease-disease similarities, disease-protein associations, chemical space networks, genome-wide association studies, and phenome-wide association studies.

With the same focus on reproducibility as Bio2RDF, OmniPath, and Pathway Commons as well as deference to software maintainability and the ease of development and inclusion of new biological data sources, we have developed a growing list of *Bio2BEL* packages, each capable of downloading, structuring, and serializing various biological data sources to BEL (**Table 2**). Each can be found in the Bio2BEL GitHub organization (https://github.com/bio2bel) as an independent open-source Python package that can readily be installed with `pip`. We have also developed and freely provided a framework (https://github.com/bio2bel/bio2bel) in the Python programming language to enable code reuse and the fast generation of additional Bio2BEL packages. Notably, the list of Bio2BEL packages includes one for OmniPath as a proof of concept that authors of other resources can implement their own Bio2BEL packages. In this article, we present the philosophy and implementation of Bio2BEL packages, a summary of past and future Bio2BEL packages, and finally, several case studies including the utility of Bio2BEL packages during curation of pathway mappings, in the analysis of cancer genome data, and for machine learning applications.

2. Implementation

Bio2BEL comprises numerous independent open-source Python packages that each enable reproducible access to a given biological data source (**Figure 1**). Each Bio2BEL package contains five components: 1) a definition of the source database or knowledge base, 2) an automated downloader for the data, 3) a parser for the data, 4) a storage

[5] http://obofoundry.org/ontology/ro
[6] https://psicquic.github.io/PSIMITAB.html

and querying system for the data, and 5) a protocol for serializing the data to BEL (**Figure 2**). In this section, we outline the components of a Bio2BEL package and their implementation details.

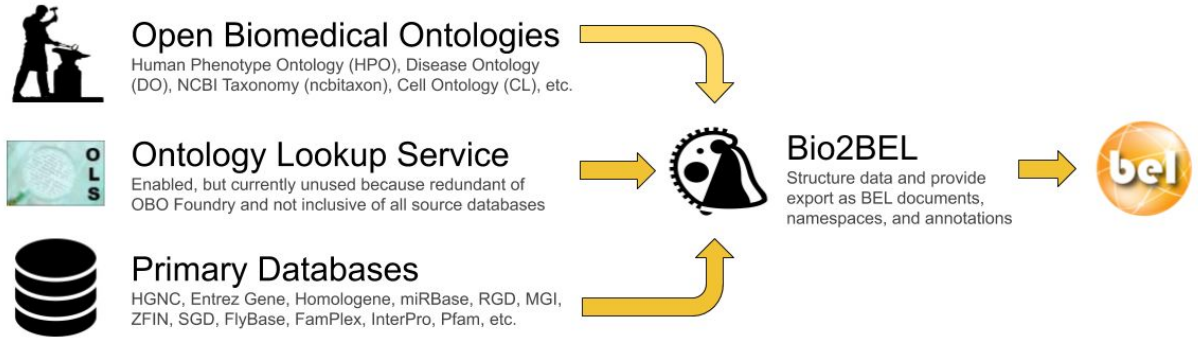


Figure 1: Though their main focus is on generating BEL documents, some Bio2BEL repositories have secondary goals of generating the BEL namespace and annotation files necessary to support manual curation. Most rely on primary databases, but the Bio2BEL framework also includes functions for generating them from standard Open Biomedical Ontology documents, or through the EBI Ontology Lookup Service (Cote *et al*., 2006). Logos adapted from http://obofoundry.org, https://www.ebi.ac.uk/ols, and https://openbel.org.

2.1. Components of a Bio2BEL Package

As this section outlines the core components and philosophy of a Bio2BEL package, it illustrates the tasks and thought process of a scientific software developer as they implement a new Bio2BEL package.

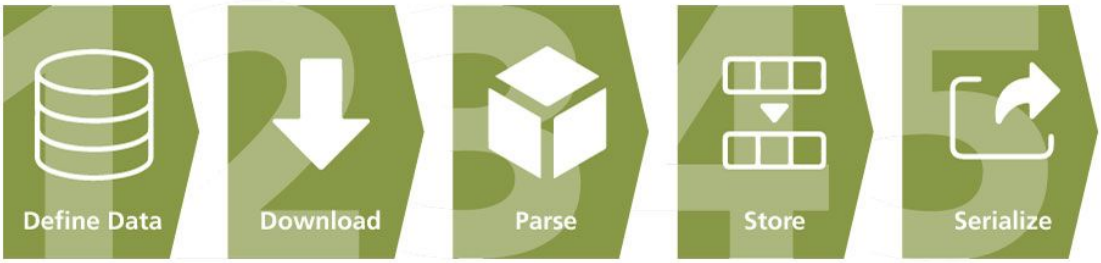


Figure 2: A graphical overview of the sequentially ordered components of a Bio2BEL package. These components correspond to the philosophy that reproducibility and accessibility can ultimately lead to the democratization of the usage of prior biological knowledge.

1. Definition of Data. The first step in generating a Bio2BEL package is to understand the source data. This requires determining if the data are publicly accessible, if they are versioned (and how the location changes with versions), and if they are available under a permissive license. Bio2BEL packages do not contain data themselves and only refer to the locations of the original data sources. For those that are versioned, providers commonly generate symlinks to the most recent version (e.g., InterPro; ftp://ftp.ebi.ac.uk/pub/databases/interpro). These characteristics help minimize licensing issues while enabling the resulting packages to update their content without changing code. Then, the developer implements custom code that makes the appropriate interpretations to convert the source data to BEL. Below, three types of data that can be readily integrated in BEL are described along with accompanying **Table 1**.

Data Source	Data Source	Example BEL statement(s)	Description

Type			
Taxonomies, Hierarchies, and Ontologies	MeSH	`path(X) isA path(Y)`	Pathology X is a subtype of pathology Y
Taxonomies, Hierarchies, and Ontologies	Complex Portal	`p(X) partOf complex(Y)`	Protein X is a member of complex Y.
Taxonomies, Hierarchies, and Ontologies	GO	`bp(X) partOf bp(Y)`	Biological process X is a sub-process of Y.
Tabular and Relational Data	PubChem, ChEMBL	`a(X) directlyDecreases act(p(Y), ma(kin))`	Compound X inhibits kinase Y.

Tabular and Relational Data	ADEPTUS	`path(X) positiveCorrelation r(Y)`
		`path(X) negativeCorrelation r(Y)`
		`path(X) causeNoChange r(Y)`
			Gene Y has been observed to either be up-regulated, down-regulated, or unregulated in patients with pathology X

Graphs	Menche *et al.*	`path(X) association path(Y)`	Pathology X is statically similar to pathology Y on the basis of gene overlap as defined by Menche *et al.* (2015).

Table 1. Example BEL statements generated by several different types of data sources

I. Taxonomies, Hierarchies, and Ontologies

The Medical Subject Headings (MeSH; Rogers, 1963) multi-hierarchy can be converted to BEL by generating an *isA* relationship between each MeSH descriptor and all of its corresponding parents in the associated MeSH tree. Nomenclatures like the Complex Portal (Meldal *et al*., (2015) also define *partOf* relations between protein complexes and their substituents. The multi-hierarchy in Gene Ontology (GO; Carbon *et al.*, 2017) can be converted similarly, which contains both *isA* relations and *partOf* relations.

II. Tabular and Relational Data

Enzyme inhibitors from ChEMBL and PubChem can be encoded like *a(X) directlyDecreases act(p(Y), ma(kin))*, and disease-specific differential gene expression can be encoded like *path(X) positiveCorrelation r(Y)* or *path(X) negativeCorrelation r(Y),* or *path(X) causeNoChange r(Y)* depending on whether the gene's expression is up-regulated, down-regulated, or not regulated, respectively. Further, BEL relationships can be extended include metadata (i.e., annotations) describing their quantitative aspects. For example, IC_{50}, EC_{50}, or other kinetic assay measurements as well as provenance and biological contextual information (e.g., original publication, cell line, assay type) can be included with the enzyme inhibition relationships from ChEMBL. Similarly, the $\log_2$ fold change and *p*-values can be included with relationships about differential gene expression.

III. Graphs

Wet-laboratory experimentation can be used to generate networks of directly observed phenomena (e.g., protein-protein interaction networks) and indirectly observed phenomena (e.g., gene co-expression networks). Graphs are often distributed as tabular data to include additional information about their constituent nodes and edges and there is often overlap with the previous data type describing tabular and relational data. *In silico* experimentation can also be used to derive edges from experimental data sets or even other graphs. For instance, bipartite graphs can be projected to homogeneous graphs consisting of a single entity and edge type as suggested by Sun *et al.* (2014).

Menche *et al.* (2015) used this strategy and computed a homogenous graph of disease-disease associations from a bipartite graph of diseases and their associated genes.

2. Downloader. The Bio2BEL framework follows a functional programming paradigm to provide an abstraction of the acquisition of data over common internet protocols like HTTP, HTTPS, and FTP. With only the URL of the data set as an input, Bio2BEL generates a download function that wraps Python's built-in *urllib* module and a simple caching mechanism in the local filesystem that avoids unnecessary network usage and duplication of potentially large files. However, some data sources, such as DrugBank (Wishart *et al.*, 2018), are not available without authentication and cannot make use of this abstraction. In those cases, developers can substitute the standard code provided in the Bio2BEL framework with custom implementations. We have taken this route for several of the packages presented in the Results section of this paper for repositories including DrugBank and MSigDB (Liberzon *et al.*, 2015).

3. Parser. There are several common file formats used by biological data sources (e.g., CSV, TSV, XML, RDF, JSON, KGML[7], Stockholm[8], OBO[9], OWL[10]). Data may also (and sometimes only) be accessible through public application programming interfaces (APIs) such as the data from KEGG (Kanehisa *et al.,* 2017), Reactome (Fabregat *et al.*, 2018), and BioThings (Xin *et al*., 2016). Alternatively, data may be available through software packages usage such the Affymetrix R package (Gautier *et al.*, 2004) and HaploReg (Ward and Kellis, 2012). After each Bio2BEL package's downloader generates a local copy of the data, the developer can either use one of the pre-defined parser functions from the Bio2BEL framework or implement a custom parser. For the most simple formats (i.e., CSV and TSV), the Bio2BEL framework automatically generates a parser that uses the *pandas* package (McKinney, 2010; https://github.com/pandas-dev/pandas). Formats like XML, JSON, and Stockholm have corresponding parsers built into the Python language or standard biology-focused packages, but the information contained within often needs custom logic for restructuring such as in the case of KGML, BioPAX, or PSI MI-XML [11]. The remaining custom formats all require custom parsers and logic. We have already implemented Bio2BEL that used CSV and TSV data (e.g., InterPro, ExCAPE-DB), XML (e.g., DrugBank), RDF (e.g., WikiPathways), JSON and KGML (e.g., KEGG), Stockholm (e.g., miRBase), and OBO and OWL (e.g., GO, DOID).

In the case of tabular data, the developer has the opportunity to annotate the column headers and their corresponding data types, which are not always included in the data and may be sought from various readme files or by exploring the corresponding website. Further, the contained data might be more useful after normalization or augmentation with information from other biological data sources. Because some databases provide identifiers with redundant information, such as the duplication of the namespace in the identifier, they must be normalized. For example, each identifier in the Disease Ontology (Schriml *et al*., 2018) is prefixed by its namespace, DOID, as can be seen in the Compact URI for the entry for restless legs syndrome, DOID:DOID:0050425. In the corresponding Bio2BEL DOID package, as well as those for others (e.g., HGNC, Gene Ontology) we normalized these identifiers to remove the redundant information. Because the main Entrez Gene database does not contain crucial information for genes, such as their chromosomal coordinates in various genomic builds, we augmented the data in the Bio2BEL Entrez package for each gene with information from RefSeq so that the genomic positions and corresponding genome build for each gene were readily accessible. Additionally, several databases that reference genes only use their HGNC gene symbols and not stable identifiers, and therefore require this additional normalization step.

[7] https://www.kegg.jp/kegg/xml/
[8] http://sonnhammer.sbc.su.se/Stockholm.html
[9] https://owlcollab.github.io/oboformat/doc/GO.format.obo-1_4.html
[10] https://www.w3.org/OWL/
[11] http://psidev.info/mif

4. Storage. Though this step may be considered optional after parsing the data, it is helpful for future reuse to choose a database type and develop a schema with which the data can be stored. Often, relational databases that can be queried with SQL are an appropriate choice. The Bio2BEL framework provides a full harness for generating an object-relational mapping (ORM) using the SQLAlchemy (https://www.sqlalchemy.org) Python package that handles generation of the SQL schema and storage of the data in a SQL database. Corresponding entity-relation diagrams can be found in the supplementary data repository at https://github.com/bio2bel/bio2bel-manuscript-supplement. While all Bio2BEL packages have, until now, used SQL databases with the SQLAlchemy ORM, there exists alternatives such as graph databases built on RDF or property graphs like Neo4J or OrientDB with a corresponding object-graph mapper that have been successfully employed in downstream applications using biological knowledge graphs (Himmelstein *et al.*, 2017; Saqi *et al.*, 2018).

5. Serializer. The final aspect of a Bio2BEL package is either to serialize the parsed data as BEL or to export the accompanying database as BEL. Entities in the SQL database that correspond to nodes and edges in BEL graphs can be converted by extending their respective ORM classes with Python functions using the internal domain-specific language provided by PyBEL (Hoyt *et al.*, 2018a). It can then be output in several formats provided by PyBEL and its growing ecosystem of plugins as well as it shields Bio2BEL packages from changes to the BEL language. Additionally, some Bio2BEL packages wrap standard nomenclature resources such as HGNC (Yates *et al*., 2017) and are able to generate BEL namespace files that are a necessary in both manual and automated curation of content in BEL (**Figure 2**). This step is deeply connected with the prior step related to the definition of the data.

2.2. Implementation Details

The Bio2BEL framework and Bio2BEL packages are implemented in Python with accessibility and readability in mind. The framework provides an abstract class *bio2bel.Manager* whose functionality all Bio2BEL packages must completely implement. Using these definitions, the framework automatically generates a uniform command line interface (CLI) that includes functions for populating the database, clearing the database, reloading data from the source, generating a web application with a view over the contents of the database, and serializing to BEL.

The Bio2BEL framework and Bio2BEL packages use flake8 (https://github.com/PyCQA/flake8) to enforce code quality, a setup.cfg file to describe the package, 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 host documentation, pytest (https://github.com/pytest-dev/pytest) as a testing framework, coverage (https://github.com/nedbat/coveragepy) and Codecov (https://codecov.io) to monitor testing coverage, and Travis-CI (https://travis-ci.com) as a continuous integration service. Further, we provide a template for Cookiecutter (https://github.com/audreyr/cookiecutter) at https://github.com/bio2bel/bio2bel-cookiecutter such that the structure of new packages can be quickly generated containing all of the configuration for each of these tools.

2.3. Implications of the Bio2BEL Philosophy

Because all Bio2BEL packages are uniform in their implementation and CLI usage, it is trivial to provide a Dockerfile and Docker-Compose configuration for quick deployments. In the future, we plan to automatically generate RESTful APIs, which may be more useful to deploy internally than to use publicly available ones due to constraints like rate-limits. Because all Bio2BEL packages are independent, they avoid two major problems of monolithic codebases: they are more robust to breakages or failures in a single package and they can be installed as needed, which is pertinent as the data sources become larger, more heterogeneous, and more complex.

Further, Bio2BEL packages can be generated by any group, and registered with the Bio2BEL framework using Python entry points (https://packaging.python.org/specifications/entry-points) that can be defined in the installation

configuration. While the Cookiecutter template allows new developers to quickly generate a package with the correct format, a full tutorial for implementing a uniform Bio2BEL package can be found at https://bio2bel.readthedocs.io/en/latest/tutorial.html.

3. Results

After describing the Bio2BEL framework and the requirements for implementing new Bio2BEL packages, we present a list of the independent Bio2BEL packages that we have already implemented in **Table 2**. We note that several of the data sources have already been included in other meta-databases like Pathway Commons and Bio2RDF, but we have chosen to implement the Bio2BEL packages using the source data rather than deriving results from these databases to provide a complementary resource for those familiar with and interested in using BEL. This choice also reduces dependencies on other projects that may not be maintained and protects against data loss during multiple conversions.

While there are thousands of high quality databases available, including a high percentage that do not fit into the schemata defined by Pathway Commons, Bio2RDF, or other meta-databases that are more appropriate for BEL, we have prioritized them as they become have become relevant for our specific use-cases, but also are open to suggestions via the issue tracker on https://github.com/bio2bel/bio2bel/issues. Below, we present four of these use cases.

Name	Description	Terms	Relations
adeptus	Disease-specific differential gene expression		4,943
chebi	Chemical multi-hierarchy	138,863	
compath	Pathway-pathway equivalences and hierarchies		1,795
ddr	Disease-disease relationships		2,997
drugbank	Drug-target interactions	11,292	25,199
entrez	Genes and orthologies	388,986	
excape	Chemical-target interactions		70,850,163
expasy	Enzyme classification and membership	6,718	243,914
famplex	Protein family and complex hierarchy		4,462
flybase	Drosophila gene nomenclature and orthologies	245,565	
go	Biological process multi-hierarchy	45,018	92,905
hgnc	Human gene nomenclature and orthologies to mouse and rat	42,741	38,360
hgncgenefamily	Human gene-gene family memberships	1,157	23,881
hippie	Protein-protein physical interactions		340,629
homologene	Gene ortholog group memberships	30,492	131,558
hsdn	Disease-symptom associations		10,246
interpro	Protein-family and protein-domain memberships	36,524	34,611
kegg	Protein-pathway memberships	330	30,346
mgi	Mouse genome nomenclature	300,499	
mirbase	MicroRNA nomenclature	38,589	
mirtarbase	miRNA-target interactions		366,110
msig	Gene-gene set memberships	17,810	2,443,391

pfam	Protein-protein family and protein family-clan memberships	17,929	
phewascatalog	Gene-disease relationships		364,667
phosphosite	Post-translational modifications		553,716
reactome	Protein-pathway and chemical-pathway memberships	23,621	137,768
rgd	Rat gene nomenclature	44,970	
sider	Drugs' side effects and indications		339,742
wikipathways	Protein-pathway memberships	513	22,115

Table 2. A non-exhaustive list of biological data sources already available as Bio2BEL packages

3.1. Mapping Concepts Between Pathway Databases with ComPath

Pathway databases have become one of the most frequently used biological data sources in the interpretation of high-throughput *-omics* experiments. Connecting pathway knowledge across the hundreds of databases developed in recent decades would not only provide a more comprehensive overview of the underlying biology they represent, but would also enable performing identical analyses on different databases. However, integrative approaches which combine databases lack the equivalence mappings between similar concepts and qualifiers that are necessary to compare between analyses run using one or another database. There are several reasons that explain the lack of mappings between databases, such as the absence of a common pathway nomenclature, differences in databases' scopes, and the lack of clear pathway boundary definitions. Furthermore, generating high quality mappings requires a significant amount of manual effort since curators must individually investigate each pair of pathways and assess whether the pair comprises related or similar pathways occurring in the same biological context.

Three Bio2BEL packages were implemented for major pathway databases (i.e., KEGG, Reactome, and WikiPathways) and extended with tools to support the first curation of mappings between their equivalent and hierarchically related pathways during the ComPath project (Domingo-Fernández *et al.*, 2018). Each were used to store and harmonize the data underlying ComPath and its accompanying web curation interface (https://compath.scai.fraunhofer.de). Though the databases of the Bio2BEL packages are detached from the ComPath web application, they can be used to integrate additional biological data sources into ComPath in the future and also to regularly update their content over time (Wadi *et al.*, 2016); thus, facilitating the revisitation and reevaluation of the mappings.

3.2. Harmonizing Pathway Databases into a Common Schema with PathMe

The most direct and effective approach in addressing issues of interoperability of pathway databases is in the transformation of various database formats into a common schema. Although this approach has been exemplified by previously mentioned databases (e.g., OmniPath, Pathway Commons, and *graphite*; Sales *et al*., 2018), there have been several limitations which have impeded a complete harmonization of pathways from distinct biological data sources. Specifically, this requires: the harmonization of biological entities to identifiers from a common nomenclature (e.g., Entrez Gene or HGNC for human genes, ChEBI or PubChem for chemicals, etc.), the normalization of biological relationships, and an underlying format which serves as the unifying schema. However, a complete harmonization risks the loss of some information in the transformation process. For instance, pathway knowledge representations can span across several scales, such as molecular events, cellular processes, and phenotypes, which various formats accommodate for in varying degrees. While existing biological data sources can address certain aspects of these steps, addressing all of these steps would enable the complete interoperability of pathway databases. Accordingly, the PathMe software was designed to harmonize pathway databases into BEL as a common representation schema with Bio2BEL at its core (Domingo-Fernández *et al.*, 2019).

The selection of BEL lies in its flexibility to incorporate a wide range of biological entities from standardized nomenclatures and their relationships, all on a multi-modal scale. The transformation of various pathway formats into BEL through PathMe is facilitated by the Bio2BEL framework by allowing for the automation of the acquisition of the biological data sources which can change frequently. By integrating PathMe and Bio2BEL, any number of pathway resources included in the latter can be transformed into BEL. In doing so, users can enrich pathway knowledge by leveraging multiple, equivalent pathway representations from the various biological data sources included in Bio2BEL and analyze their own networks alongside canonical pathway ones. In a later publication, we plan to demonstrate the utility of combining Bio2BEL packages to produce an integrative pathway resource. Similarly to the recent comparison of pathway activity measurement tools by Lim *et al*. (2018), we will benchmark the performance of each of these resources both individually and combined on functional pathway enrichment and classification tasks applied to cancer genome and patient data.

3.3. Applications of Network Representation Learning with BioKEEN

The integration of numerous biological databases into a common schema gives rise to large, rich, heterogeneous knowledge graphs to which a variety of statistical and machine learning methodologies can be applied. One family of approaches, network representation learning (NRL), has been shown to be useful for clustering, entity disambiguation, and link prediction tasks (Nickel *et al.*, 2016). As new machine learning models are published for accomplishing these tasks, several implementations using the currently popular machine learning frameworks TensorFlow (Abadi *et al.* 2016) and PyTorch (Paszke *et al.*, 2017) provide reference implementations.

We developed BioKEEN as an extension to the previously developed NRL package, PyKEEN, to enable it to directly acquire and preprocess BEL knowledge graphs, namely those generated by Bio2BEL (Ali *et al*., 2018). One of the original goals of PyKEEN was to democratize NRL methods by facilitating those less familiar with the relevant mathematics and programming backgrounds to apply and evaluate them. We have continued this philosophy with BioKEEN to allow scientists to specify the Bio2BEL packages they would like to include in their analysis that are either hosted on PyPI, GitHub, or already installed as custom local packages. The usage of Bio2BEL allows scientists using NRL as a component of a more complex analytical pipeline to have the ability to not only re-run analyses in a reproducible manner, but also make use of the ability to acquire updated data when it becomes available.

Along with our previous publication, we provided several demonstrations including the prediction of novel protein-protein interactions using a model trained with the BioKEEN package for the Human Integrated Protein-Protein Interaction rEference (HIPPIE; Alanis-Lobato *et al.*, 2017), the prediction of pathway mappings using ComPath, and the prediction of disease-symptom associations using the Bio2BEL package for the HSDN (Zhou *et al.*, 2014) provided by Himmelstein *et al.* (2017) with Rephetio (https://het.io). Later, we plan to apply BioKEEN to combinations of Bio2BEL repositories to support other biologically relevant link prediction tasks such as drug repositioning.

3.4. Interoperability with Other Projects

The Integrated Network and Dynamical Reasoning Assembler (INDRA; Gyori *et al.*, 2017) integrates several databases including those covering physical interactions (e.g., BioGrid; Chatr-Aryamontri *et al.*, 2017), signaling (e.g., SIGNOR; Perfetto *et al*. 2016), curated drug targets (e.g., HMS LINCS small molecule target relationship database; http://lincs.hms.harvard.edu), and experimental drug affinities (e.g., Target Affinity Spectrum; Moret *et al*., 2018) in order to support generation of dynamical models. Following the recent development of a converter between BEL and INDRA (Hoyt *et al.*, 2019), these biological data sources can be indirectly made available as BEL, and all Bio2BEL packages can be integrated in INDRA.

Similarly, we are collaborating with the researchers developing OmniPath to structure their data acquisition pipelines as a Bio2BEL package, which is currently under development. Notably, OmniPath encompasses several biological data sources related to protein-protein interactions, transcriptional regulation, post-translational modifications, ligand-receptor interactions, and protein complexes, and others. This resource is complementary to content already available through Bio2BEL, providing a more comprehensive integration of the extensive publicly available biological data sources.

4. Conclusions

While the development of Bio2BEL has addressed the lack of defined schemata, data standardization, annotation of entities with classes, and application of controlled vocabularies to relations in numerous biological databases by converting them to BEL, several considerations remain. The approaches taken by Bio2RDF, Pathway Commons, and now Bio2BEL can be categorized as *data warehousing*. An alternative strategy, *data federation*, attempts to combine disparate biological data sources using SPARQL endpoints (e.g., DisGeNet-RDF (Queralt-Rosinach *et al.*, 2016), UniProt (Redaschi *et al.*, 2009), EBI (Jupp *et al.*, 2014)), RESTful APIs (e.g., BioServices (Cokelaer *et al.*, 2013), BioThings, Orange Bioinformatics (Curk *et al.*, 2005)), and more recently, GraphQL (https://graphql.org). Bio2BEL does not directly address data federation, but other aspects of the BEL ecosystem such as BEL Commons (Hoyt *et al.*, 2018b) have exposed RESTful APIs for manipulating BEL that might also be useful for GraphQL. However, the several attempts[12,13,14] at converting BEL to RDF have suffered from relatively low adoption; and while a conversion to RDF enables querying with SPARQL, BEL lacks a dedicated query language that can leverage the rich aspects of its statements beyond their subjects, predicates, and objects.

Finally, it remains that like any format, consumers of BEL must make their own transformations appropriate for their scientific applications. We are not discouraged by this fact, and believe that Bio2BEL is a step towards enabling more computational scientists easy access to a larger portion of the wealth of available structured biological knowledge resources.

5. Availability and requirements

Project Name: Bio2BEL

Project Home Page: https://github.com/bio2bel

Operating System(s): Platform independent

Programming Language: Python 3

License: MIT License

Declarations

Ethics Approval and Consent to Participate

Not applicable

Consent for Publication

Not applicable

[12] https://wiki.openbel.org/display/OBP/BEL+RDF+Model
[13] https://github.com/OpenBEL/bel2rdf
[14] https://github.com/cthoyt/cx-rdf

Availability of Data and Materials

Each Bio2BEL package is listed on https://github.com/bio2bel and automatically acquires relevant data from their respective original biological data sources.

Competing Interests

The authors declare that they have no competing interests.

Funding

This work was partially 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.

This work was also partially supported by the Fraunhofer Society's MAVO program.

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

Authors' Contributions

CTH conceived and designed the study. CTH, DDF, and SM drafted the manuscript. MHA acquired funding and reviewed the manuscript. All authors performed data curation and developed computational pipelines for extraction, transformation, and loading of various biological data sources. All authors have read and approved the final manuscript.

Acknowledgements

We would like to thank the curators and maintainers of the several databases we have used, without whom none of this work would be possible.

References

1. Abadi, M., *et al.* (2016). TensorFlow: A System for Large-scale Machine Learning. In *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation* (pp. 265–283). Berkeley, CA, USA: USENIX Association. Retrieved from http://dl.acm.org/citation.cfm?id=3026877.3026899
2. Alanis-Lobato, G., Andrade-Navarro, M. A., & Schaefer, M. H. (2017). HIPPIE v2.0: enhancing meaningfulness and reliability of protein–protein interaction networks. *Nucleic Acids Research*, *45*(D1), D408–D414. https://doi.org/10.1093/nar/gkw985
3. Ali, M., *et al*. (2018). BioKEEN: A library for learning and evaluating biological knowledge graph embeddings, 1–5. https://doi.org/10.1101/475202
4. Alocci, D., *et al*. (2015). Property Graph vs RDF triple store: A comparison on glycan substructure search. *PLoS ONE*, *10*(12), 1–17. https://doi.org/10.1371/journal.pone.0144578
5. Belleau, F., Nolin, M.-A., Tourigny, N., Rigault, P., & Morissette, J. (2008). Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. *Journal of Biomedical Informatics*, *41*(5), 706–716. https://doi.org/10.1016/j.jbi.2008.03.004

6. Carbon, S., *et al.* (2017). Expansion of the gene ontology knowledgebase and resources: The gene ontology consortium. *Nucleic Acids Research*, *45*(D1), D331–D338. https://doi.org/10.1093/nar/gkw1108
7. Cerami, E. G., *et al.* (2011). Pathway commons, a web resource for biological pathway data. *Nucleic acids research*. *39*(Suppl. 1), D685–D690. https://doi.org/10.1093/nar/gkq1039.
8. Chatr-Aryamontri, A., *et al.* (2017). The BioGRID interaction database: 2017 update. *Nucleic Acids Research*, *45*(D1), D369–D379. https://doi.org/10.1093/nar/gkw1102
9. Chen, B., *et al.* (2010). Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data. *BMC Bioinformatics*, *11*, 255.
10. Cokelaer, T., Pultz, D., Harder, L. M., Serra-Musach, J., & Saez-Rodriguez, J. (2013). BioServices: a common Python package to access biological Web Services programmatically. *Bioinformatics*, *29*(24), 3241–3242. https://doi.org/10.1093/bioinformatics/btt547
11. Cote, R., *et al*. (2006). The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries. *BMC Bioinformatics*, *7*, 1–7. https://doi.org/10.1186/1471-2105-7-97
12. Courtot, M., *et al.* (2011). Controlled vocabularies and semantics in systems biology. *Molecular Systems Biology*, *7*(543). https://doi.org/10.1038/msb.2011.77
13. Curk, T., *et al*. (2005). Microarray data mining with visual programming. *Bioinformatics*, *21*(3), 396–398. https://doi.org/10.1093/bioinformatics/bth474
14. Davidson, S. B., Overton, C., & Buneman, P. (1995). Challenges in integrating biological data sources. *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology*, *2*(4), 557–572. https://doi.org/10.1089/cmb.1995.2.557
15. Demir, E., *et al*. (2010). The BioPAX community standard for pathway data sharing. *Nature Biotechnology*, *28*(12), 1308–1308. https://doi.org/10.1038/nbt1210-1308c
16. Domingo-Fernández, D., *et al.* (2018). ComPath: An ecosystem for exploring, analyzing, and curating mappings across pathway databases. *npj Systems Biology and Applications*, *4*(1):43. https://doi.org/10.1038/s41540-018-0078-8.
1. Domingo-Fernandez, D., Mubeen, S., Marin-Llao, J., Hoyt, C., & Hofmann-Apitius, M. (2019). PathMe: Merging and exploring mechanistic pathway knowledge. *BMC Bioinformatics,* 20:243. https://doi.org/10.1186/s12859-019-2863-9.
17. Emon, M. A. E. K., Kodamullil, A. T., Karki, R., Younesi, E., & Hofmann-Apitius, M. (2017). Using Drugs as Molecular Probes: A Computational Chemical Biology Approach in Neurodegenerative Diseases. *Journal of Alzheimer's Disease*, *56*(2), 677–686. https://doi.org/10.3233/JAD-160222
18. Fabregat, A., *et al.* (2018). The Reactome Pathway Knowledgebase. *Nucleic Acids Research*, *46*(D1), D649–D655. https://doi.org/10.1093/nar/gkx1132
19. Fan, X.-N., Zhang, S.-W., Zhang, S.-Y., Zhu, K., & Lu, S. (2019). Prediction of lncRNA-disease associations by integrating diverse heterogeneous information sources with RWR algorithm and positive pointwise mutual information. *BMC Bioinformatics*, *20*(1), 87. https://doi.org/10.1186/s12859-019-2675-y
20. Gautier, L., Cope, L., Bolstad, B. M., & Irizarry, R. A. (2004). Affy - Analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, *20*(3), 307–315. https://doi.org/10.1093/bioinformatics/btg405
21. Gyori, B. M., *et al*. (2017). From word models to executable models of signaling networks using automated assembly. *Molecular Systems Biology*, *13*(11), 954. https://doi.org/10.15252/msb.20177651
22. Himmelstein, D. S., Lizee, A., Hessler, C., Brueggeman, L., Chen, S. L., Hadley, D., … Baranzini, S. E. (2017). Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *ELife*, *6*. https://doi.org/10.7554/eLife.26726
23. Hoyt, C. T., *et al.* (2019). Re-curation and Rational Enrichment of Knowledge Graphs in Biological Expression Language. *Database : The Journal of Biological Databases and Curation*, baz068. https://doi.org/10.1093/database/baz068
24. Hoyt, C. T., Konotopez, A., & Ebeling, C. (2018a). PyBEL: a computational framework for Biological Expression Language. *Bioinformatics*, *34*(4), 703–704. https://doi.org/10.1093/bioinformatics/btx660

25. Hoyt, C. T., Domingo-Fernández, D., & Hofmann-Apitius, M. (2018b). BEL Commons: an environment for exploration and analysis of networks encoded in Biological Expression Language. *Database : The Journal of Biological Databases and Curation*, *2018*(3), 1–11. https://doi.org/10.1093/database/bay126
26. Irin, A. K., Tom Kodamullil, A., Gündel, M., & Hofmann-Apitius, M. (2015). Computational Modelling Approaches on Epigenetic Factors in Neurodegenerative and Autoimmune Diseases and Their Mechanistic Analysis. *Journal of Immunology Research*, *2015*, 1–10. https://doi.org/10.1155/2015/737168
27. Iyappan, A., *et al.* (2014). NeuroRDF : Semantic Data Integration Strategies for Modeling Neurodegenerative Diseases. *Proceedings of the 6th International Symposium on Semantic Mining in Biomedicine (SMBM2014)*, (January 2016), 11–18.
28. Iyappan, A., *et al*. (2017). Neuroimaging Feature Terminology: A Controlled Terminology for the Annotation of Brain Imaging Features. *Journal of Alzheimer's Disease*, *59*(4), 1153–1169. https://doi.org/10.3233/jad-161148
29. Jupp, S., *et al*. (2014). The EBI RDF platform: Linked open data for the life sciences. *Bioinformatics*, *30*(9), 1338–1339. https://doi.org/10.1093/bioinformatics/btt765
30. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., & Morishima, K. (2017). KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, *45*(D1), D353–D361. https://doi.org/10.1093/nar/gkw1092
31. Laibe, C., & Le Novère, N. (2007). MIRIAM Resources: tools to generate and resolve robust cross-references in Systems Biology. *BMC Systems Biology*, *1*, 58. https://doi.org/10.1186/1752-0509-1-58
32. Liberzon, A., *et al.* (2015). The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Systems*, *1*(6), 417–425. https://doi.org/10.1016/J.CELS.2015.12.004
33. Lim, S., Lee, S., Jung, I., Rhee, S., & Kim, S. (2018). Comprehensive and critical evaluation of individualized pathway activity measurement tools on pan-cancer data. *Briefings in bioinformatics*. https://doi.org/10.1093/bib/bby097
34. McKinney, W. (2010). Data Structures for Statistical Computing in Python. In S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 51–56).
35. Meldal, B. H. M., *et al*. (2015). The complex portal - An encyclopaedia of macromolecular complexes. *Nucleic Acids Research*, *43*(D1), D479–D484. https://doi.org/10.1093/nar/gku975
36. Menche, J., *et al*. (2015). Disease networks. Uncovering disease-disease relationships through the incomplete interactome. *Science (New York, N.Y.)*, *347*(6224), 1257601. https://doi.org/10.1126/science.1257601
37. Moret, N., *et al*. (2018). Cheminformatics tools for analyzing and designing optimized small molecule libraries. *BioRxiv*, (617), 358978. https://doi.org/10.1101/358978
38. Naz, M., Kodamullil, A. T., & Hofmann-Apitius, M. (2016). Reasoning over genetic variance information in cause-and-effect models of neurodegenerative diseases. *Briefings in Bioinformatics*, *17*(3), 505–16. https://doi.org/10.1093/bib/bbv063
39. Nickel, M., Murphy, K., Tresp, V., & Gabrilovich, E. (2016). A Review of Relational Machine Learning for Knowledge Graphs. *Proceedings of the IEEE*, *104*(1), 11–33. https://doi.org/10.1109/JPROC.2015.2483592
40. Paszke, A., Chanan, G., Lin, Z., Gross, S., Yang, E., Antiga, L., & Devito, Z. (2017). Automatic differentiation in PyTorch. *31st Conference on Neural Information Processing Systems*, (Nips), 1–4. https://doi.org/10.1017/CBO9781107707221.009
41. Perfetto, L., *et al*. (2016). SIGNOR: A database of causal relationships between biological entities. *Nucleic Acids Research*, *44*(D1), D548–D554. https://doi.org/10.1093/nar/gkv1048
42. Queralt-Rosinach, N., Piñero, J., Bravo, À., Sanz, F., & Furlong, L. I. (2016). DisGeNET-RDF: harnessing the innovative power of the Semantic Web to explore the genetic basis of diseases. *Bioinformatics*, *32*(14), 2236–2238. https://doi.org/10.1093/bioinformatics/btw214

43. Redaschi, N., & Consortium, U. (2009). UniProt in RDF: Tackling Data Integration and Distributed Annotation with the Semantic Web. *Nature Precedings*. https://doi.org/10.1038/npre.2009.3193.1
44. Rogers, F. B. (1963). Medical subject headings. Bulletin of the Medical Library Association, 51, 114–6. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/13982385
45. Sales, G., *et al.* (2018). *meta*Graphite - a new layer of pathway annotation to get metabolite networks, *Bioinformatics*, bty719. https://doi.org/10.1093/bioinformatics/bty719.
46. Saqi, M., Lysenko, A., Guo, Y.-K., Tsunoda, T., & Auffray, C. (2018). Navigating the disease landscape: knowledge representations for contextualizing molecular signatures. *Briefings in Bioinformatics*, (May), 1–15. https://doi.org/10.1093/bib/bby025
47. Schriml, L. M., *et al*. (2018). Human Disease Ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Research*, 1–8. https://doi.org/10.1093/nar/gky1032
48. Slater, T. (2014). Recent advances in modeling languages for pathway maps and computable biological networks. *Drug Discovery Today*, *19*(2), 193–198. https://doi.org/10.1016/j.drudis.2013.12.011
49. Stelzer, G., *et al.* (2016). The GeneCards suite: From gene data mining to disease genome sequence analyses. *Current Protocols in Bioinformatics*, *2016*(June), 1.30.1-1.30.33. https://doi.org/10.1002/cpbi.5
50. Sun, K., Pržulj, N., Buchan, N., & Larminie, C. (2014). The integrated disease network. *Integrative Biology*, *6*(11), 1069–1079. https://doi.org/10.1039/c4ib00122b
51. Szklarczyk, D., *et al.* (2015). STRING v10: Protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, *43*(D1), D447–D452. https://doi.org/10.1093/nar/gku1003
52. Türei, D., Korcsmáros, T., & Saez-Rodriguez, J. (2016). OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nature methods*, *13*(12):966. https://doi.org/10.1038/nmeth.4077.
53. van Dam, J. C. J., Schaap, P. J., Martins dos Santos, V. A. P., & Suárez-Diez, M. (2014). Integration of heterogeneous molecular networks to unravel gene-regulation in Mycobacterium tuberculosis. *BMC Systems Biology*, *8*(1), 111. https://doi.org/10.1186/s12918-014-0111-5
54. Wadi, L., *et al.* (2016). Impact of outdated gene annotations on pathway enrichment analysis. *Nature methods*, *13*(9):705. https://doi.org/10.1038/nmeth.3963.
55. Wanichthanarak, K., Fahrmann, J. F., & Grapov, D. (2015). Genomic, proteomic, and metabolomic data integration strategies. *Biomarker Insights*, *10*(Table 1), 1–6. https://doi.org/10.4137/BMI.S29511
56. Ward, L. D., & Kellis, M. (2012). HaploReg: A resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Research*, *40*(D1), 930–934. https://doi.org/10.1093/nar/gkr917
57. Warde-Farley, D., *et al.* (2010). The GeneMANIA prediction server: Biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Research*, *38*(SUPPL. 2), 214–220. https://doi.org/10.1093/nar/gkq537
58. Williams, A. J., *et al*. (2012). Open PHACTS: semantic interoperability for drug discovery. *Drug Discovery Today*, *17*(21–22), 1188–1198. https://doi.org/10.1016/j.drudis.2012.05.016
59. Wishart, D. S., *et al*. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research*, *46*(D1), D1074–D1082. https://doi.org/10.1093/nar/gkx1037
60. Xin, J., *et al*. (2016). High-performance web services for querying gene and variant annotation. *Genome Biology*, *17*(1), 91. https://doi.org/10.1186/s13059-016-0953-9
61. Yates, B., *et al*. (2017). Genenames.org: The HGNC and VGNC resources in 2017. *Nucleic Acids Research*, *45*(D1), D619–D625. https://doi.org/10.1093/nar/gkw1033
62. Zhou, X., Menche, J., Barabási, A.-L., & Sharma, A. (2014). Human symptoms-disease network. *Nature Communications*, *5*(May), 4212. https://doi.org/10.1038/ncomms5212

Conclusions

6 BEL2ABM: agent-based simulation of static models in Biological Expression Language

Introduction

Bioinformatics, 34(13), 2018, 2316–2318
doi: 10.1093/bioinformatics/bty107
Advance Access Publication Date: 23 February 2018
Applications Note

Systems biology

BEL2ABM: agent-based simulation of static models in Biological Expression Language

Michaela Gündel[1,2,*], Charles Tapley Hoyt[1,2] and Martin Hofmann-Apitius[1,2]

[1]Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Sankt Augustin 53754, Germany and [2]Department of Life Science Informatics, Bonn-Aachen International Center for IT, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn 53113, Germany

*To whom correspondence should be addressed.
Associate Editor: Bonnie Berger

Received on July 25, 2017; revised on December 9, 2017; editorial decision on February 20, 2018; accepted on February 21, 2018

Abstract

Summary: While cause-and-effect knowledge assembly models encoded in Biological Expression Language are able to support generation of mechanistic hypotheses, they are static and limited in their ability to encode temporality. Here, we present BEL2ABM, a software for producing continuous, dynamic, executable agent-based models from BEL templates.
Availability and implementation: The tool has been developed in Java and NetLogo. Code, data and documentation are available under the Apache 2.0 License at https://github.com/pybel/bel2abm.
Contact: martin.hofmann-apitius@scai.fraunhofer.de
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The ability of Biological Expression Language (BEL) to encode qualitative cause-and-effect relationships from biological systems makes it well-suited for generating mechanistic hypotheses in the context of experimental data (Catlett *et al.*, 2013). However, it generally lacks the ability to describe the temporal evolution of dynamic systems except in special cases where time can be represented discretely. For example, the progression of Alzheimer's disease (AD) is often discretized to healthy, mild cognitive impairment and full AD. While other systems biology modeling languages such as Systems Biology Markup Language can natively embed mathematical equations in order to support simulations and produce generative models (Hucka *et al.*, 2003), they lack the ability of BEL to represent multi-scale and multi-modal processes.

To the best of our knowledge, there have not been any previously published attempts to produce dynamic biological models from BEL. A simple approach to introduce dynamism would be to convert to rule-based models such as Petri nets. However, discrete dynamism is inherently limited in its expressivity. Continuous dynamism can be achieved through agent-based modeling (ABM); where a discrete number of entities, their properties and their methods of interaction are encoded and simulated in a complex system. In addition, ABMs do not require fine-granular knowledge of the reaction rates and other kinetic properties of a system that often limit the utility of mathematical models.

Here, we present BEL2ABM, a software package that transforms static BEL knowledge assemblies into continuous, dynamic, executable, ABMs.

2 Materials and methods

All physical entities in a BEL knowledge assembly are encoded as agents with properties (e.g. lifespans, ability to reproduce, etc.) described by the Human Physiology Simulation Ontology (HuPSON) (Gündel *et al.*, 2013) and behaviors derived from their relationships to other entities in the BEL knowledge assembly. Statements like `A increases B` creates a behavior where a unilateral coupling between `A` and `B` stochastically increases the number or scale of B. For example, the statement `p(A) increases kinaseActivity (p(B))` represents that when the two proteins A and B are within interaction distance, the kinase activity property

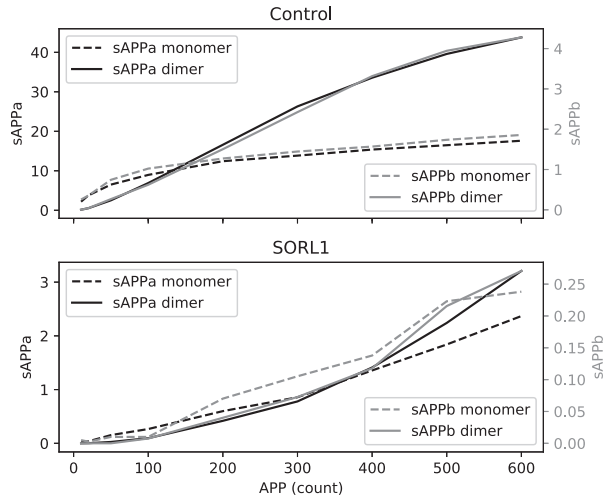


Fig. 1. Plotted are average initial rates of the production of sAPPa, sAPPb and their respective homodimers in the BEL2ABM simulation based on the amyloid beta cascade with varying initial number of APP agents. The sigmoidal curves observed in the control correspond to the cooperativity of the allosteric secretase dimers of sAPPa and sAPPb. Perturbation with SORL1 inhibits oligomerization, shown by the loss of sigmoidal shape, as well as causing significant decrease in the production of each of sAPPa, sAPPb and their respective homodimers. Settings alpha secretases: 10; alpha secretase dimers: 10; beta secretases: 1; beta secretase dimers: 1; SORL1: 3; APP binding sites of allosteric enzymes: 2; binding strengths: 95%; high lifespans so few molecule dies during experiment; 400 replicate runs at each APP concentration

of B is increased. Biological processes are translated to procedures which have effects determined by HuPSON over a large physical area within the ABM simulation. Finally, additional information is encoded about each entity type that is available from context-specific annotations in BEL (e.g. cellular locations) and HuPSON. A complete schema for converting BEL to ABM properties and behaviors can be found in the Supplementary Material.

Temporality is introduced with NetLogo (https://ccl.northwestern.edu/netlogo) which updates agents based on simulation parameters and their internal properties at small discrete time points to simulate continuous time. It provides an environment where users can adjust modeling parameters (i.e. molecule numbers, interaction distances, movement speed, etc.) and produce replicates.

Finally, Spartan (Alden *et al.*, 2014) is used to perform consistency analysis, evaluate the results and establish the minimal number of replicates needed. For many applications of ABMs, the number of each type of entity at each time point is of great interest. Using the replicates from Spartan, the numbers at each time point are averaged in order to provide more robust results.

3 Case study

The processing of amyloid beta precursor-protein (APP) in the amyloid cascade has been highly implicated in AD. Recent experimental evidence has shown not only that the α-secretase and β-secretase enzymes act cooperatively as allosteric homodimers in the cleavage of APP, but this process is also inhibited by sortilin related receptor 1 (SORL1) through the blocking APP oligomerization (Schmidt *et al.*, 2012).

As a case study, we encoded the relevant entities, processes and relations from the amyloid beta cascade in BEL in order to assess the ability of BEL2ABM to produce an ABM that replicates the sigmoidal patterns of enzyme cooperativity observed in experimental observations and captured in ordinary differential equations (ODE) shown in Figure 11 of Schmidt *et al.*

While the resulting ABM provides a wealth of information about each of the entities involved in the system, Figure 1 presents the most relevant measurements representing the respective initial rates of production of sAPPa and sAPPb as a function of the initial amount of APP present that can be compared to Figure 11 of Schmidt *et al.* Both the sAPPa and sAPPb curves adhere to the sigmoidal patterns observed in vitro and described by the ODE from Schmidt *et al.* Additionally, the presence of SORL1 in the ABM also reproduced the behavior of significantly decreasing the production rates of sAPPa and sAPPb, solely based on the knowledge encoded in BEL of the agents, their properties and their behaviors.

After investigating the model's robustness to parameter settings using Spartan, we concluded that this model was sufficiently robust to both simulation stochastic noise (with a minimum of 400 replicates) and to parameter value change. The case of the initial α-secretase number (at $t = 0$) showed medium to medium-high effect over a large range (10–600) of initial APP entities; whereas changes in α-secretase binding strength showed only a small effect for low, and medium effect for high APP numbers. The underlying BEL document, BEL resources, NetLogo settings and results can all be found at https://github.com/pybel/bel2abm.

4 Discussion

Because BEL2ABM produces inherently qualitative models, several constraints must be considered during their evaluation. The magnitude of the results cannot be directly compared to experimental results or mathematical models such as the ODE system provided by Schmidt *et al.* because time and space are only artificially incorporated during simulation. Thus, we only expect to observe similar behavioral patterns of a real biological system.

NetLogo and other common simulation environments allow users to modify various simulation parameters in order to improve the adherence of an ABM to experimental data. While this allows users to the benefit of exploring based on their intuition, systematic optimization becomes a combinatorial problem for larger and more complex systems. BEL2ABM includes some semi-automatic parameter optimization methods and can be theoretically run with an optimization procedure like a grid search, but future work will include developing and encoding more biologically-driven optimization procedures in an ontology, like HuPSON, that can be leveraged to more automatically build relevant models. Further, the burden of choosing the most relevant and informative knowledge assemblies for BEL2ABM may be eased by the hypothesis generation procedures in upcoming BEL frameworks like PyBEL (Hoyt *et al.*, 2018).

With these restrictions in mind, we have shown that it is possible to dynamize a static knowledge assembly model, enable a user to qualitatively reproduce the behavior a biological system, and modify model parameters in order to make further investigations.

Acknowledgements

This work has been supported by the B-IT Foundation.

Conflict of Interest: none declared.

References

Alden,K.J. *et al.* (2014) Applying spartan to understand parameter uncertainty in simulations. *R. J.*, **6**, 63–80.

Catlett,N.L. *et al*. (2013) Reverse causal reasoning: applying qualitative causal knowledge to the interpretation of high-throughput data. *BMC Bioinformatics*, **14**, 340.

Gündel,M. *et al*. (2013) HuPSON: the human physiology simulation ontology. *J. Biomed. Seman*., **4**, 35.

Hoyt,C.T. *et al*. (2018) PyBEL: a computational framework for Biological Expression Language. *Bioinformatics*, **34**, 703–704.

Hucka,M. *et al*. (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics (Oxford, England)*, **19**, 524–531.

Schmidt,V. *et al*. (2012) Quantitative modelling of amyloidogenic processing and its influence by SORLA in Alzheimer's disease. *EMBO J*., **31**, 187–200.

Conclusions

7 GuiltyTargets: prioritization of novel therapeutic targets with deep network tepresentation learning

Introduction

GuiltyTargets: Prioritization of Novel Therapeutic Targets with Deep Network Representation Learning

Özlem Muslu[1,2], Charles Tapley Hoyt[1,2], Martin Hofmann-Apitius[1,2], Holger Fröhlich[2,3,*]

1 Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Schloss Birlinghoven, Sankt Augustin, 53754, Germany
2 Bonn-Aachen International Center for Information Technology (B-IT), University of Bonn, 53115, Bonn, Germany
3 UCB Biosciences GmbH, Alfred-Nobel Str. 10, 40789 Monheim, Germany

*Corresponding author: frohlich@bit.uni-bonn.de

Abstract

The majority of clinical trial failures are caused by low efficacy of investigated drugs, often due to a poor choice of target protein. Computational prioritization approaches aim to support target selection by ranking candidate targets in the context of a given disease. We propose a novel target prioritization approach, GuiltyTargets, which relies on deep network representation learning of a genome-wide protein-protein interaction network annotated with disease-specific differential gene expression and uses positive-unlabeled machine learning for candidate ranking. We evaluated our approach on six diseases of different types (cancer, metabolic, neurodegenerative) within a 10 times repeated 5-fold stratified cross-validation and achieved AUROC values between 0.92 - 0.94, significantly outperforming a previous approach, which relies on manually engineered topological features. Moreover, we showed that GuiltyTargets allows for target repositioning across related disease areas. Applying GuiltyTargets to Alzheimer's disease resulted into a number of highly ranked candidates that are currently discussed as targets in the literature. Interestingly, one (COMT) is also the target of an approved drug (Tolcapone) for Parkinson's disease, highlighting the potential for target repositioning of our method.

Availability: The GuiltyTargets Python package is available on PyPI and all code used for analysis can be found under the MIT License at https://github.com/GuiltyTargets.

Author summary

Many drug candidates fail in clinical trials due to low efficacy. One of the reasons is the choice of the wrong target protein, i.e. perturbation of the protein does not effectively modulate the disease phenotype on a molecular level. In consequence many patients do not demonstrate a clear response to the drug candidate. Traditionally, targets are selected based on evidence from the literature and follow-up experiments. However, this process is very labor intensive and often biased by subjective choices. Computational tools could help a more rational and unbiased choice of target proteins and thus increase the chance of drug discovery programs. In this work we propose a novel machine learning based method for target candidate ranking. The method (GuiltyTargets) captures properties of known targets to learn a ranking of candidates. GuiltyTargets compares favorably against existing machine learning based target prioritization methods and allowed us to propose novel targets for Alzheimer's disease.

Introduction

Drug discovery is a time consuming, expensive and complicated process [1–4]. Many drug candidates fail in clinical studies due to low efficacy, mainly because of wrong target choice [5–7]. Traditionally, scientists identified targets by searching through the relevant literature, following clues from mRNA and protein expression, integrating expression data with pathway analyses, experimenting with knockout mice, investigating somatic mutations, gene fusions, and copy number variations, and using the accumulated knowledge from multiple experimental studies to generate a hypothesis on how a molecule might work as a target [2, 8, 9]. However, manually interpreting many data sources is prone to biased identification of targets as it limits the potential to use all available and helpful data. By computationally integrating multiple biological data sources to analyze prior knowledge, it should be possible to make target identification process faster, less biased, and more informed. Computational target prioritization approaches thus aim for improving target identification process by ranking proteins based on their likelihood of being targets in the context of a specific disease [10–17]. Most of them integrate biological networks with other data sources to prioritize targets for infectious diseases [11–13], cancers [14–16] or neurodegenerative diseases [17].

In addition, machine learning methods have been used to prioritize drug targets. For example, Emig *et al.* proposed an approach, in which for each candidate target a number of different network topological features are combined with proximity to differentially expressed genes in a particular disease of interest [10]. All features are subsequently combined into a logistic regression model, which allows for a ranking of candidate targets. The authors successfully tested their approach with 30 different diseases. Another example is the method by Ferrero *et al.*, which uses features provided by the Open Targets database [18] and combines them into one ranking score using Support Vector Machines [19].

In this paper we propose a novel approach to prioritize targets using a combination of unsupervised network representation learning, namely the recently proposed Gat2Vec method [20], and logistic regression. More specifically, our method, GuiltyTargets, first maps a genome-wide protein-protein interaction network annotated with differential gene expression information into an Euclidean space using Gat2Vec. In that space, we then use positive-unlabeled (PU) machine learning [21–24] to learn a ranking of candidate targets. To the best of our knowledge, network representation learning as a data driven approach to implicitly learn relevant topological features from a network structure has not been used for target prioritization so far. The proposed approach is compared to the approach from Emig *et al.* [10] for six diseases, demonstrating its superior ranking performance. For the example of Alzheimer's disease (AD), in-depth analysis shows that GuiltyTargets can be used to reposition known targets from other neurological indications.

Results

GuiltyTargets: A New Approach for Deep Network Representation Learning Based Target Prioritization

Our newly developed GuiltyTargets method can be summarized as follows (Figure 1): First, a genome-wide PPI network is compiled and annotated with discretized information about differential gene expression within a given disease context (-1 = underexpressed; 0 = no significant change; 1 = overexpressed). Next, the attributed network is embedded into an Euclidean space using Gat2Vec. Following a PU learning scheme known disease specific protein targets are assigned positive labels, and the

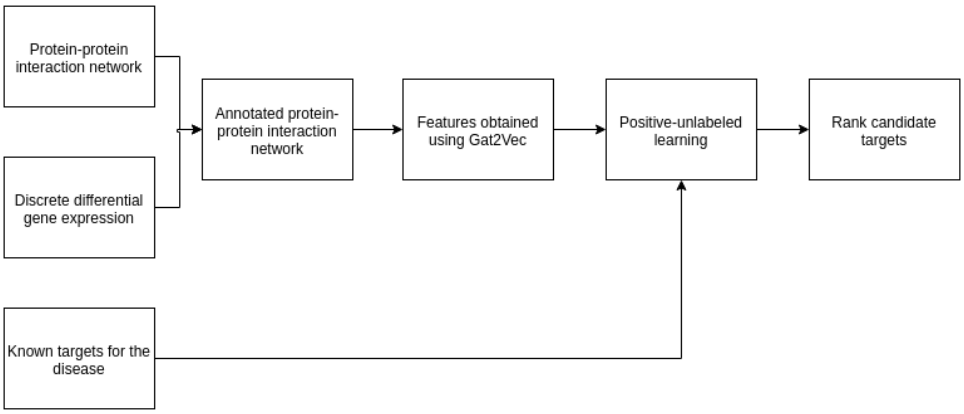


Fig 1. GuiltyTargets pipeline. First, a protein-protein interaction network is annotated with discrete differential gene expression information (up-regulated, down-regulated, not differentially expressed). 128 features are extracted from this annotated network using Gat2Vec. These features are input to a logistic regression algorithm where known targets are labeled as positive and the remaining as negative. The likelihoods that are calculated by the classifier is then used for ranking.

remaining proteins are regarded as pseudo-negatives to train a classifier that ranks a candidate protein according its similarity to known targets for the given disease. More details about GuiltyTargets are described in the Methods section of this paper.

Validation Data

We performed target prioritization analyses for six different diseases using corresponding gene expression data for acute myeloid leukemia, hepatocellular carcinoma, idiopathic pulmonary fibrosis, liver cirrhosis, multiple sclerosis and AD. The choice was made based on the following criteria: First, five of these diseases have also been evaluated in the publication by Emig *et al.* [10], which we used for comparison here. Second, the number of available known targets for each disease was expected to be relatively high for a statistically meaningful validation. Finally, we added AD to investigate the applicability of our approach to a highly challenging disease, in which so far most attempts to establish new drugs have failed [25]. Details about about used data, including pre-processing, are described in Materials and Methods Section. Notably, for AD we investigated RNASeq data from different cohorts (MSBB [26], MayoRNASeq [27], ROSMAP [28]). To investigate the prediction performance of GuiltyTargets we employed two protein-protein interaction networks (STRING [29], HIPPIE [30]), two target databases (Open Targets [18], Therapeutic Target Database [31]) and different cutoffs to discretize differential gene expression via log_2 fold change thresholds (0.5, 1.0, 1.5) while requiring a false discovery rate of less than 5%.

GuiltyTargets Outperforms Existing Method

The performance of our approach and the method by Emig *et al.* were compared within a 10 times repeated 5-fold cross validation scheme with the area under ROC curve (AUROC) as the evaluation criterion. This assessed the ability of each method to rank in an independent test set a true known target higher than an unknown protein. For this purpose, the approach employed by Emig *et al.* was re-implemented using the same PPI network resources and target databases as for GuiltyTargets.

Disease	**Emig *et al.* (original)**	**Emig *et al.***	**GuiltyTargets**
Acute myeloid leukemia	0.8195	0.8356 ± 0.0001	0.9277 ± 0.0002
Alzheimer's disease	-	0.6235 ± 0.0010	0.9418 ± 0.0004
Hepatocellular carcinoma	0.8019	0.7314 ± 0.0002	0.9384 ± 0.0001
Idiopathic pulmonary fibrosis	0.8826	0.8306 ± 0.0018	0.9263 ± 0.0004
Liver cirrhosis	0.6747	0.5338 ± 0.0019	0.9464 ± 0.0002
Multiple sclerosis	0.7151	0.6755 ± 0.0002	0.9412 ± 0.0003

Table 1. Ranking performance of GuiltyTargets compared to the method by Emig *et al.* in terms of cross-validated AUROC (standard error). The table shows the results when using Open Targets as resource for drug targets, STRING (using the default confidence threshold) as PPI network and declaring differential gene expression based on log2 fold change cutoff of 1.5, which is agreement to Emig *et al.* The column "Reported" shows the AUROC values reported by Emig *et al.* and the column "Implemented" shows the AUROC values obtained by the reimplementation of their method. The row "Alzheimer's Disease" refers to the MayoRNAseq data from temporal cortex, which among all tested AD gene expression datasets showed genes with an absolute log2 fold change larger than 1.5.

Results shown in Tables 1, S1 and S2 demonstrate a dramatic performance increase of up to 40% by GuiltyTargets compared to the method by Emig *et al.*. Notably, AUROC values found by our re-implementation of were not identical (but typically close) to the ones reported in the original paper. This was likely due to the fact that not the same PPI network and target database resources have been used. More specifically, Emig *et al.* employed the commercial MetaBase™database, whereas here we only rely on public resources.

We employed an ANOVA to assess the statistical significances of our findings. The ANOVA model was built separately for each investigated scenario (PPI network, target database, log2 fold change cutoff) with three factors: method, dataset and an interaction term between method and dataset. The ANOVA F-test confirmed a highly significant improvement of GuiltyTargets compared to the method by Emig *et al.* for each disease and scenario ($p < 0.001$ after Holm's correction for multiple testing, see Figure 2). As expected there was a highly significant dataset dependency of AUROC performance in every case ($p < 2.2e-16$). Post-hoc analysis using Tukey's multiple comparisons of means revealed that, on average, GuiltyTargets outperformed the approach by Emig *et al.* by 14.8% AUROC.

In-Depth Analysis of Influence Factors on GuiltyTargets Performance

We wanted to better understand the dependency of the performance of GuiltyTargets on the different tested influence factors, which we had varied individually in our cross-validation analysis:

- PPI network (STRING, HIPPIE), including different confidence level thresholds
- Target database (Open Targets, Therapeutic Targets Database)
- Thresholds for declaring differential gene expression

For this purpose we fitted a two-way ANOVA model with interaction term (influence factor, dataset, interaction term between both) and then performed a Tukey post-hoc analysis. Table 2 demonstrates that the employed PPI network is the most relevant influence factor for GuiltyTargets: Using STRING significantly increased the AUROC compared to using HIPPIE by 9%. On the other hand, the chosen log2 fold change

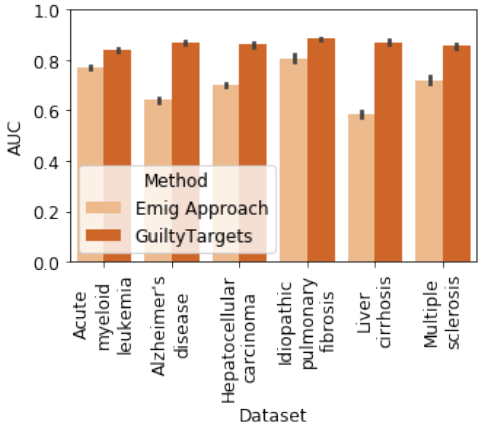


Fig 2. frohlich: Please change legend: Comparison of GuiltyTargets vs approach by Emig *et al.*: The barplots show AUC values averaged over all possible hyper-parameter choices (log2 fold change cutoff, target database, PPI network, PPI network confidence cutoff).

Influence factor	**Comparison**	**Difference**	**p-value**
PPI network	STRING vs. HIPPIE	0.09	1.15E-14
PPI confidence threshold	default vs 0.63	0.0318	1.15E-14
Target database	Open Targets vs. TTD	0.0527	2.23E-10
	0.5 vs 1.0	-0.0016	0.01369
Log2 fold change cutoff	0.5 vs 1.5	0.0007	0.46799
	1.0 vs 1.5	0.0023	0.00026

Table 2. In-depth analysis of different influence factors on the performance of GuiltyTargets. The table shows the result of the Tukey ANOVA post-hoc analysis of different comparisons. The difference in AUROC is shown in column 3 together with the corresponding p-value in column 4.

threshold had almost no influence, i.e. GuiltyTargets is highly robust against this parameter. A more conservative confidence threshold for the STRING network yielded a drop in prediction performance by 3%. Both findings together may be explained by a comparably strong influence of the network topology for GuiltyTargets, which is leveraged by Gat2Vec. An obvious question is therefore, in how far GuiltyTargets is affected at all by gene expression data or whether our method is purely topology based. We thus compared the performance of our method across the three tested AD gene expression datasets (MSBB, MayoRNASeq, ROSMAP), confirming a significant influence of the actually used dataset on the AUROC for this particular disease ($p = 3.09E-09$, ANOVA F-test). Hence, gene expression data does have a clear effect on GuiltyTargets.

The use of the Open Targets versus the Therapeutic Target Database significantly increased the ranking performance of GuiltyTargets by 5%, hence underlining the relevance of a larger number of known targets for learning the ranking model in the embedded network space.

GuiltyTargets Learns from Known Targets

We tested, whether the good performance of GuiltyTargets was dependent on known targets or whether also with a random set of proteins a similar performance could have been achieved. For this purpose we trained GuiltyTargets for each disease with 100 randomly drawn sets of targets of the same size as the actual ones, which we

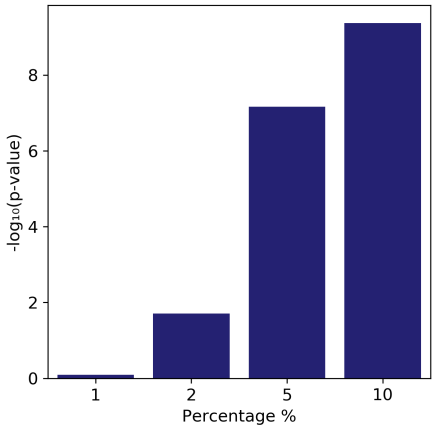


Fig 3. Target repositioning potential of GuiltyTargets: The barplot shows the result of a hypergeometric test conducted on the top p% of a ranked list of candidate proteins when looking for overrepresentation of known AD targets. GuiltyTargets was trained without any known AD targets here.

incorrectly labeled as "targets". Prediction performance was evaluated using the same cross-validation procedure as before. Table S3 confirms that the AUROC for random proteins drops to about 50%, i.e. chance level. Hence, GuiltyTargets indeed learns properties of known targets.

GuiltyTargets Allows for Target Repositioning Across Related Diseases

There is the question whether GuiltyTargets could transfer properties learned from known targets in one disease area into another one, hence allowing for repositioning of targets. To address that question we trained GuiltyTargets with all known targets of neurodegenerative diseases obtained from Open Targets, while excluding known AD targets. We then ran a hypergeometric test on the resulting prioritization to see if known AD targets where statistically overrepresented at the top of the list. The results were significant when at least 2% of the top candidates were considered (Figure 3). This shows that GuiltyTargets could help for repositioning targets across related disease areas.

Case Study: GuiltyTargets Predicts New Candidate Targets for Alzheimer's Disease

Despite of 180 therapeutic targets listed in the Open Targets database, the AD field urgently requires new and more effective medications that either prevent, mitigate, or reverse its progression. This is particularly true, because the vast majority of drugs under development fail in clinical trials [25]. We here picked out AD as a test case for GuiltyTargets to prioritize new target candidates. We used post-mortem gene expression data from brain tissue from the ROSMAP study and combined it with STRING based PPI network and Open Targets as a resource for known targets. ROSMAP data was chosen because of its comparably large number of samples (495 AD patients and 438 controls). Table 3 shows the top 0.1% of a ranked list of novel candidate targets (likelihood score > 0.4) obtained with GuiltyTargets. According to the TTD [31] and

Entrez identifier	HGNC symbol	Likelihood score	Class	Known drugs / druggable
1143	CHRNB4	0.7	Nicotinic acetylcholine receptor	SIB-1553A (Alzheimer, discontinued in Phase 2)
3708	ITPR1	0.689	IP3 receptor	yes
2742	GLRA2	0.619	Ligand gated chloride channel	yes
1312	COMT	0.587	Catechol-O-Methyltransferase	Tolcapone (Parkinson)
2898	GRIK2	0.587	Ionotropic glutamate receptor	yes
1132	CHRM4	0.586	Muscarinic acetylcholine receptor	yes
89832	CHRFAM7A	0.557	Nicotinic acetylcholine receptor	yes
3363	HTR7	0.532	Serotonin receptor	JNJ-18038683(Major depressive disorder) ATI-9242 (Schizophrenia, discontinued in Phase 2)
3777	KCNK3	0.523	Potassium channel	yes
2741	GLRA1	0.484	Glycine receptor	D-Serine(Parkinson's disease, Phase 4)
1136	CHRNA3	0.461	Nicotinic acetylcholine receptor	yes
3269	HRH1	0.451	Histamine receptor	Doxepin(Depression) Doxylamine(Anxiety disorder) Propiomazine(Insomnia, Anxiety disorder)] Pyrilamine Maleate(headache)
6543	SLC8A2	0.448	Solute carrier	
2911	GRM1	0.447	Metabotropic glutamate receptor	PF-1913539 (Alzheimer's disease, discontinued in Phase 3) A-841720 (Pain, preclinical) AZD8529 (Scizophrenia, discontinued in Phase 2)
2913	GRM3	0.445	Metabotropic glutamate receptor	BCI-632 (Alzheimer disease, Major depressive disorder, Phase 1) Pomaglumetad (Schizophrenia, Phase 1) LY-544344 (Anxiety disorder, Discontinued in Phase 3) LY354740 (Anxiety disorder, Discontinued in Phase 2) R-1578 (Mood disorder, Discontinued in Phase 2) RO-4995819 (Major depressive disorder, Discontinued in Phase 2)
3361	HTR5A	0.436	Serotonin receptor	yes
8001	GLRA3	0.412	Ligand gated chloride channel	yes
1360	CPB1	0.411	Protaminase	yes
9456	HOMER1	0.407	Neuronal immediate-early gene	

Table 3. Target prioritization for AD using ROSMAP gene expression data. This list shows candidate proteins above a likelihood score threshold of 0.4. The last column shows either known drugs (including indications) against the respective target or the classification as "druggable" using the information from DGIdb [32] and TTD [31].

DGIdb [32] databases, all but two candidates are druggable, thus they could be used as targets for drugs using the current drug development methods.

Many of the candidate targets are receptors, namely four acetylcholine receptors (three nicotinic, one muscarinic), and three glutamate receptors (two metabotropic, one ionotropic), in agreement with the observation that receptors constitute a large portion of known AD targets for small molecule drugs [33]. The remaining candidates were identified as ion channels. The top candidate (CHRNB4) is the target of the compound SIB-1553A, which has been tested in a phase 2 clinical trial for AD, but discontinued (source: Therapeutic Target Database). Out of the other top candidates we found CHRFAM7A, GRM1, GRM3, ITPR1, HTR7, and COMT particularly interesting: CHRFAM7A is an alpha-7 nicotinic cholinergic receptor subunit interacting with amyloid β, whose aggregates (i.e., plaques) are one of the hallmarks of AD [34]. CHRFAM7A may promote neuronal survival and function, and subunits are expressed by astrocytes participating in synaptic communication [35]. GRM1 is the target of the compound PF-1913539, which has been discontinued in a phase 3 AD trial [31]. GRM3 (mGlu3) is found in astrocytes as well as neuronal cells, and have been observed to have neuroprotective properties. Its agonists and positive allosteric modulators were reported to be potentially helpful for AD treatment [36]. Glial mGlu3 receptors regulate the production of neurotrophic factors such as nerve growth factor, brain-derived neurotrophic factor and glial-derived neurotrophic factor [36]. BCI-632, a compound that targets GRM3, is currently tested in a phase 1 AD trial [31]. ITPR1, an intracellular Ca 2+ channel, mediates calcium release from the endoplasmic reticulum, triggering apoptosis, and its deletion has been linked to spinocerebellar ataxia type 15, a neurodegenerative disease [37, 38]. Single nucleotide polymorphisms (SNPs) rs73310256 in HTR7 [39] and rs4680 in COMT have been associated with AD [40]. COMT is currently discussed as a target for AD [41]. Finally, COMT is the target of the anti-Parkinson drug Tolcapone (source: TTD), supporting our previous finding that GuiltyTargets can re-propose targets from related diseases.

Visualization of interactions between known and candidate targets (Figure 4) revealed that the network of known and proposed targets has a higher than expected interaction rate (PPI enrichment $p < 1.0E - 16$, calculated using STRING web interface). Further-

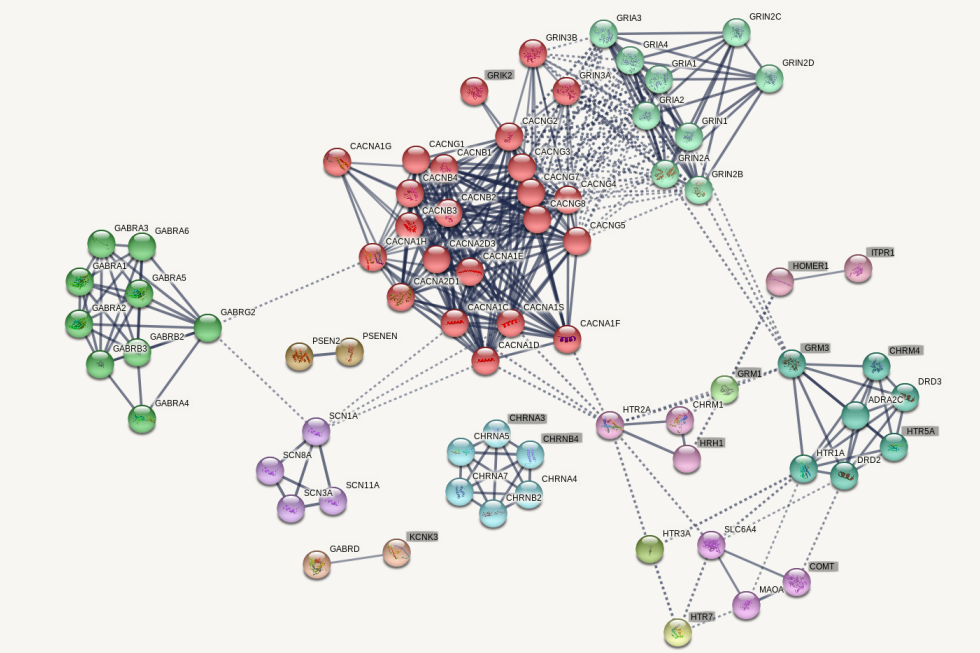


Fig 4. Interactions between known and candidate targets, with confidence scores higher than 0.7. Clusters of nodes were calculated using MCL clustering [42] with inflation parameter of 3.4 and the nodes were colored based on the clusters they are in. The transparency of the links shows the confidence score of the interaction. If a node has some known or predicted 3D structure, it is filled with a structure image. Highlighted nodes show the proposed candidates, whereas the rest show the known targets. Image generated using STRING [29].

more, the candidate targets were observed to reside on the borders of this interaction network. These two observations lead to the hypothesis that targeting the proposed candidates would propagate through the network, influencing disease-related proteins indirectly. Generating drugs that target multiple proteins in this list might be effective, if these candidates were to be considered as different entry points to the disease module.

Discussion

We presented a novel network representation learning based approach for target prioritization, GuiltyTargets. Our approach uses a protein-protein interaction network, a differential gene expression profile and a list of known targets to prioritizes proteins as targets for a particular disease. We showed that GuiltyTargets is highly robust and significantly outperforms the method by Emig *et al.* in terms of ranking performance. As demonstrated by our validation studies, it is applicable to various types of diseases, including cancers, metabolic and neurodegenerative diseases. We demonstrated that GuiltyTargets can be used to repurpose existing targets from a different (but related) disease area. Application of GuiltyTargets to AD showed that several of the highest ranked candidates are indeed proposed in the literature for AD, and three of them have been targeted by candidate AD drugs. Moreover, our case study once more demonstrated the possibility to repurpose existing targets from related disease areas with our methods, e.g. COMT as target of the drug Tolcapone in Parkinson's Disease.

GuiltyTargets as well as other machine learning based target prioritization methods

(including the one by Emig *et al.*) learn properties of known targets to rank candidate proteins. Hence, these approaches rely on available information about targets, which is often incomplete and noisy [43]. Because of their dependency on available data machine learning approaches typically have difficulties to propose candidates that point towards completely novel disease biology. Despite these limitations GuiltyTargets showed promising results for target prioritization, including our case study for AD. Hence, we see GuiltyTargets as a promising tool to support the decision process in the context of target identification in pharmaceutical research.

Materials and Methods

Information Sources

Three types of information sources were used for target prioritization:

1. differential gene expression profiles between diseased and healthy subjects
2. protein-protein interaction networks (PPI network)
3. disease-specific target annotation

In the following we provide more detailed information about these data.

Gene Expression Data

Gene expression data for acute myleoid leukemia, hepatocellular carcinoma, idiopathic pulmonary fibrosis, liver cirrhosis and multiple sclerosis was obtained from Gene Expression Omnibus (GEO) [44], and differential gene expression was assessed via GEO2R [45], Biobase [46], GEOquery [47] and limma [48] using multiple testing correction via the false discovery rate [49] (see: Supplementary Table 1).

For AD, RNASeq data from the AM-PAD Knowledge Portal (AM-PAD) was used [50]. In particular, MSBB, ROSMAP, and MayoRNASeq studies were utilized. Differential gene expression was assessed by applying DESeq2 to the normalized RNAseq data for each brain region (Table S4).

Protein-Protein Interaction Networks

As PPI networks, HIPPIE v2.0 [30] and STRING v10.5 [29] were used since both of these networks are created by combining multiple sources of PPIs and provide confidence scores. HIPPIE and STRING differ in the type of interactions they contain (Table S5): HIPPIE relies on physical protein-protein interactions, whereas STRING captures more broadly functional interactions. Hence, STRING has a much larger size than HIPPIE. The analyses on this paper only included the interactions between human proteins. STRING locus identifiers were mapped to Entrez identifies using the mappings provided by STRING.

Target Databases

Information about known targets were obtained from two databases: The Therapeutic Target Database (TTD) [31] and Open Targets [18] (see: Table S6). Target identifiers in TTD database were mapped to UniProt identifiers using the conversion file provided by TTD. These identifiers were then mapped to Entrez gene IDs using R packages AnnotationDBI [51] and org.Hs.eg.db [52]. In addition to TTD, known protein targets were retrieved from Open Targets. HGNC symbols were converted to Entrez identifiers using R packages AnnotationDBI [51] and org.Hs.eg.db [52].

GuiltyTargets

Deep Network Representation Learning

GuiltyTargets relies on a deep representation learning of an annotated PPI network via Gat2Vec, where node attributes represent discretized gene expression log fold changes (see Results part). In a first step, two separate graphs, the structural and the attribute graph, are constructed from the original labeled PPI network, where the structural graph corresponds to the PPI network, and the attribute graph is a bipartite graph between protein nodes and discretized log fold changes. Afterwards Gat2Vec retrieves for each vertex its structural context through random walks of a predefined length. The exact parameters that we used to run Gat2Vec are given in Table S7. The result of each random walk is a sequence of vertices and node attributes, respectively. These sequences are subsequently embedded into an Euclidean space using a SkipGram neural network, which is an essential part of the well known Word2Vec method [53].

Target Candidate Ranking

The features obtained from annotated PPI networks and the disease-specific target annotations were used to train a logistic regression classifier. Following a PU learning scheme known targets were assigned positive labels, and the remaining proteins were treated as if they were negatives. It is essential to note that proteins that are not known as targets in a particular disease of interest could in fact be targets in another disease context, and finding them is the primary goal target prioritization approaches. For the implementation, the LogisticRegression class from linear_model module and OneVsRestClassifier class from multiclass module in Python library scikit-learn were used [54].

Funding

This work was supported by Fraunhofer-Gesellschaft.

Acknowledgements

We would like to thank Daniel Domingo-Fernández for his valuable help in interpreting the ranked candidate list for Alzheimer's disease.

The results published here are in part based on data obtained from the AMP-AD Knowledge Portal (doi 10.7303/syn2580853).

For MSBB data set, these data were generated from postmortem brain tissue collected through the Mount Sinai VA Medical Center Brain Bank and were provided by Dr. Eric Schadt from Mount Sinai School of Medicine.

For MayoRNASeq data set, study data were provided by the following sources: The Mayo Clinic Alzheimer's Disease Genetic Studies, led by Dr. Nilufer Taner and Dr. Steven G. Younkin, Mayo Clinic, Jacksonville, FL using samples from the Mayo Clinic Study of Aging, the Mayo Clinic Alzheimer's Disease Research Center, and the Mayo Clinic Brain Bank. Data collection was supported through funding by NIA grants P50 AG016574, R01 AG032990, U01 AG046139, R01 AG018023, U01 AG006576, U01 AG006786, R01 AG025711, R01 AG017216, R01 AG003949, NINDS grant R01 NS080820, CurePSP Foundation, and support from Mayo Foundation. Study data includes samples collected through the Sun Health Research Institute Brain and Body Donation Program of Sun City, Arizona. The Brain and Body Donation Program is supported by the National Institute of Neurological Disorders and Stroke (U24 NS072026 National Brain and Tissue Resource for Parkinsons Disease and Related Disorders), the National Institute on Aging

(P30 AG19610 Arizona Alzheimer's Disease Core Center), the Arizona Department of Health Services (contract 211002, Arizona Alzheimer's Research Center), the Arizona Biomedical Research Commission (contracts 4001, 0011, 05- 901 and 1001 to the Arizona Parkinson's Disease Consortium) and the Michael J. Fox Foundation for Parkinson's Research.

For ROSMAP data set, study data were provided by the Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago. Data collection was supported through funding by NIA grants P30AG10161, R01AG15819, R01AG17917, R01AG30146, R01AG36836, U01AG32984, U01AG46152, the Illinois Department of Public Health, and the Translational Genomics Research Institute.

Author contributions statement

Ö.M. and C.T.H. implemented the method and analysed data. M.H.-A. and H.F. designed the research. H.F. supervised the project. Ö.M., C.T.H. and H.F. drafted the manuscript.

Competing Interests statement

H.F. received salaries from UCB Biosciences GmbH. UCB Biosciences GmbH had no influence on the content of this work.

References

1. Csermely P, Korcsmáros T, Kiss HJ, London G, Nussinov R. Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. Pharmacology & therapeutics. 2013;138(3):333–408.

2. Gashaw I, Ellinghaus P, Sommer A, Asadullah K. What makes a good drug target? Drug discovery today. 2011;16(23-24):1037–1043.

3. Hughes JP, Rees S, Kalindjian SB, Philpott KL. Principles of early drug discovery. British journal of pharmacology. 2011;162(6):1239–1249.

4. Lotfi Shahreza M, Ghadiri N, Mousavi SR, Varshosaz J, Green JR. A review of network-based approaches to drug repositioning. Briefings in bioinformatics. 2017;.

5. Arrowsmith J. Trial watch: Phase II failures: 2008–2010; 2011.

6. Laenen G, Thorrez L, Börnigen D, Moreau Y. Finding the targets of a drug by integration of gene expression data with a protein interaction network. Molecular BioSystems. 2013;9(7):1676–1685.

7. Arrowsmith J. Phase III and submission failures: 2007-2010. Nature Reviews Drug Discovery. 2011;10(2):1–2.

8. Isik Z, Baldow C, Cannistraci CV, Schroeder M. Drug target prioritization by perturbed gene expression and network information. Scientific reports. 2015;5:17417.

9. L Moseley F, Bicknell K, S Marber M, Brooks G. The use of proteomics to identify novel therapeutic targets for the treatment of disease. The Journal of pharmacy and pharmacology. 2007;59:609–28. doi:10.1211/jpp.59.5.0001.

10. Emig D, Ivliev A, Pustovalova O, Lancashire L, Bureeva S, Nikolsky Y, et al. Drug target prediction and repositioning using an integrated network-based approach. PLoS One. 2013;8(4):e60618.

11. Doyle MA, Gasser RB, Woodcroft BJ, Hall RS, Ralph SA. Drug target prediction and prioritization: using orthology to predict essentiality in parasite genomes. BMC genomics. 2010;11(1):222.

12. Paul MS, Kaur A, Geete A, Sobhia ME. Essential gene identification and drug target prioritization in Leishmania species. Molecular BioSystems. 2014;10(5):1184–1195.

13. Gupta SK, Gross R, Dandekar T. An antibiotic target ranking and prioritization pipeline combining sequence, structure and network-based approaches exemplified for Serratia marcescens. Gene. 2016;591(1):268–278.

14. Yeh SH, Yeh HY, Soo VW. A network flow approach to predict drug targets from microarray data, disease genes and interactome network-case study on prostate cancer. Journal of clinical bioinformatics. 2012;2(1):1.

15. Vitali F, Cohen LD, Demartini A, Amato A, Eterno V, Zambelli A, et al. A network-based data integration approach to support drug repurposing and multi-target therapies in triple negative breast cancer. PloS one. 2016;11(9):e0162407.

16. Bidkhori G, Benfeitas R, Elmas E, Kararoudi MN, Arif M, Uhlen M, et al. Metabolic Network-Based Identification and Prioritization of Anticancer Targets Based on Expression Data in Hepatocellular Carcinoma. Frontiers in physiology. 2018;9.

17. Keane H, Ryan BJ, Jackson B, Whitmore A, Wade-Martins R. Protein-protein interaction networks identify targets which rescue the MPP+ cellular model of Parkinson's disease. Scientific reports. 2015;5:17004.

18. Koscielny G, An P, Carvalho-Silva D, Cham JA, Fumis L, Gasparyan R, et al. Open Targets: a platform for therapeutic target identification and validation. Nucleic acids research. 2016;45(D1):D985–D994.

19. Ferrero E, Dunham I, Sanseau P. In silico prediction of novel therapeutic targets using gene–disease association data. Journal of translational medicine. 2017;15(1):182.

20. Sheikh N, Kefato Z, Montresor A. gat2vec: representation learning for attributed graphs. Computing. 2018; p. 1–23.

21. Li XL, Liu B. Learning from positive and unlabeled examples with different data distributions. In: European Conference on Machine Learning. Springer; 2005. p. 218–229.

22. Peng L, Zhu W, Liao B, Duan Y, Chen M, Chen Y, et al. Screening drug-target interactions with positive-unlabeled learning. Scientific Reports. 2017;7(1):8087.

23. Hameed PN, Verspoor K, Kusljic S, Halgamuge S. Positive-unlabeled learning for inferring drug interactions based on heterogeneous attributes. BMC bioinformatics. 2017;18(1):140.

24. Yang P, Li XL, Mei JP, Kwoh CK, Ng SK. Positive-unlabeled learning for disease gene identification. Bioinformatics. 2012;28(20):2640–2647.

25. Mehta D, Jackson R, Paul G, Shi J, Sabbagh M. Why do trials for Alzheimer's disease drugs keep failing? A discontinued drug perspective for 2010-2015. Expert opinion on investigational drugs. 2017;26(6):735–739.

26. Wang M, Beckmann ND, Roussos P, Wang E, Zhou X, Wang Q, et al. The Mount Sinai cohort of large-scale genomic, transcriptomic and proteomic data in Alzheimer's disease. Scientific data. 2018;5:180185.

27. Allen M, Carrasquillo MM, Funk C, Heavner BD, Zou F, Younkin CS, et al. Human whole genome genotype and transcriptome data for Alzheimer's and other neurodegenerative diseases. Scientific data. 2016;3:160089.

28. Mostafavi S, Gaiteri C, Sullivan SE, White CC, Tasaki S, Xu J, et al. A molecular network of the aging human brain provides insights into the pathology and cognitive decline of Alzheimer's disease. Nature neuroscience. 2018;21(6):811.

29. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, et al. The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. Nucleic acids research. 2016; p. gkw937.

30. Alanis-Lobato G, Andrade-Navarro MA, Schaefer MH. HIPPIE v2. 0: enhancing meaningfulness and reliability of protein–protein interaction networks. Nucleic acids research. 2016; p. gkw985.

31. Li YH, Yu CY, Li XX, Zhang P, Tang J, Yang Q, et al. Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics. Nucleic acids research. 2017;46(D1):D1121–D1127.

32. Cotto KC, Wagner AH, Feng YY, Kiwala S, Coffman AC, Spies G, et al. DGIdb 3.0: a redesign and expansion of the drug–gene interaction database. Nucleic Acids Research. 2018;46(D1):D1068–D1073. doi:10.1093/nar/gkx1143.

33. Santos R, Ursu O, Gaulton A, Bento AP, Donadi RS, Bologa CG, et al. A comprehensive map of molecular drug targets. Nature reviews Drug discovery. 2017;16(1):19.

34. Murpy M, LeVine III H. Alzheimer's disease and the β-amyloid peptide. J Alzheimers Dis. 2010;19(1):311–323.

35. Dineley KT, Pandya AA, Yakel JL. Nicotinic ACh receptors as therapeutic targets in CNS disorders. Trends in pharmacological sciences. 2015;36(2):96–108.

36. Caraci F, Battaglia G, Sortino MA, Spampinato S, Molinaro G, Copani A, et al. Metabotropic glutamate receptors in neurodegeneration/neuroprotection: still a hot topic? Neurochemistry international. 2012;61(4):559–565.

37. Hara K, Shiga A, Nozaki H, Mitsui J, Takahashi Y, Ishiguro H, et al. Total deletion and a missense mutation of ITPR1 in Japanese SCA15 families. Neurology. 2008;71(8):547–551.

38. Van de Leemput J, Chandran J, Knight MA, Holtzclaw LA, Scholz S, Cookson MR, et al. Deletion at ITPR1 underlies ataxia in mice and spinocerebellar ataxia 15 in humans. PLoS genetics. 2007;3(6):e108.

39. Herold C, Hooli BV, Mullin K, Liu T, Roehr JT, Mattheisen M, et al. Family-based association analyses of imputed genotypes reveal genome-wide significant association of Alzheimer's disease with OSBPL6, PTPRG, and PDCL3. Molecular psychiatry. 2016;21(11):1608.

40. Corbo RM, Gambina G, Broggio E, Scarabino D, Scacchi R. Association study of two steroid biosynthesis genes (COMT and CYP17) with Alzheimer's disease in the Italian population. Journal of the neurological sciences. 2014;344(1-2):149–153.

41. Perkovic MN, Strac DS, Tudor L, Konjevod M, Erjavec GN, Pivac N. Catechol-O-methyltransferase, Cognition and Alzheimer's Disease. Current Alzheimer Research. 2018;15(5):408–419.

42. Dongen S. A cluster algorithm for graphs. 2000;.

43. Zakeri P, Simm J, Arany A, ElShal S, Moreau Y. Gene prioritization using Bayesian matrix factorization with genomic and phenotypic side information. Bioinformatics. 2018;34(13):i447–i456.

44. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic acids research. 2002;30(1):207–210.

45. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic acids research. 2012;41(D1):D991–D995.

46. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-throughput genomic analysis with Bioconductor. Nature methods. 2015;12(2):115.

47. Davis S, Meltzer PS. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics. 2007;23(14):1846–1847.

48. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic acids research. 2015;43(7):e47–e47.

49. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society Series B (Methodological). 1995; p. 289–300.

50. Hodes RJ, Buckholtz N. Accelerating medicines partnership: Alzheimer's disease (AMP-AD) knowledge portal aids Alzheimer's drug discovery through open data sharing; 2016.

51. Carlson M, Falcon S, Pages H, Li N. AnnotationDbi: Annotation Database Interface. R package version;1(0).

52. Carlson M, Falcon S, Pages H, Li N. org. Hs. eg. db: Genome wide annotation for Human. R package version 33. 2013;.

53. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:13013781. 2013;.

54. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2825–2830.

Conclusions

8 A systematic approach for identifying shared mechanisms in epilepsy and its comorbidities

Introduction

Database, 2018, 1–9
doi: 10.1093/database/bay050
Original article

Original article

A systematic approach for identifying shared mechanisms in epilepsy and its comorbidities

Charles Tapley Hoyt[1,2,*,†], Daniel Domingo-Fernández[1,2,†], Nora Balzer[2], Anka Güldenpfennig[1] and Martin Hofmann-Apitius[1,2]

[1]Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing, Konrad-Adenauer-Straße, Schloss Birlinghoven, 53754 Sankt Augustin, Germany and [2]Department of Life Science Informatics, Bonn-Aachen International Center for IT, Rheinische Friedrich-Wilhelms-Universität Bonn, Endenicher Allee 19C, Bonn 53113, Germany

*Corresponding author: Tel: +49 2241 14-2268; Fax: +49 2241 14-2656; Email: charles.hoyt@scai.fraunhofer.de

[†]These authors contributed equally to this work.

Citation details: Hoyt, C.T., Domingo-Fernández, D., Balzer, N. *et al.* A systematic approach for identifying shared mechanisms in epilepsy and its comorbiditie. *Database* (2018) Vol. 2018: article ID bay050; doi:10.1093/database/bay050

Received 9 February 2018; Revised 23 April 2018; Accepted 26 April 2018

Abstract

Cross-sectional epidemiological studies have shown that the incidence of several nervous system diseases is more frequent in epilepsy patients than in the general population. Some comorbidities [e.g. Alzheimer's disease (AD) and Parkinson's disease] are also risk factors for the development of seizures; suggesting they may share pathophysiological mechanisms with epilepsy. A literature-based approach was used to identify gene overlap between epilepsy and its comorbidities as a proxy for a shared genetic basis for disease, or genetic pleiotropy, as a first effort to identify shared mechanisms. While the results identified neurological disorders as the group of diseases with the highest gene overlap, this analysis was insufficient for identifying putative common mechanisms shared across epilepsy and its comorbidities. This motivated the use of a dedicated literature mining and knowledge assembly approach in which a cause-and-effect model of epilepsy was captured with Biological Expression Language. After enriching the knowledge assembly with information surrounding epilepsy, its risk factors, its comorbidities, and anti-epileptic drugs, a novel comparative mechanism enrichment approach was used to propose several downstream effectors (including the GABA receptor, GABAergic pathways, etc.) that could explain the therapeutic effects carbamazepine in both the contexts of epilepsy and AD. We have made the Epilepsy Knowledge Assembly available at https://www.scai.fraunhofer.de/content/dam/scai/de/downloads/bioinformatik/epilepsy.bel and queryable through NeuroMMSig at http://neurommsig.scai.fraunhofer.de. The source code used for analysis and tutorials for reproduction are available on GitHub at https://github.com/cthoyt/epicom.

Introduction

Seizures are transient occurrences of signs and symptoms due to abnormal or excessive neuronal activity in the brain (1). Classically, their underlying causes were thought to be the primary drivers for increasing mortality in epilepsy. While epilepsy has been classically studied as a disorder of the brain characterized by an enduring pre-disposition to epileptic seizures, it is no longer considered a condition in which seizures are the only concern (2). Epilepsy is also associated with several comorbidities, including Alzheimer's disease (AD), Parkinson's disease (PD), other nervous system diseases and psychiatric disorders (3–5), due to a variety of genetic, biological and environmental factors (6).

The prevalence of migraine in epilepsy patients (under 64 years old) is 5.71% in contrast to 3.47% in the general population (7). The mechanistic understanding of epilepsy and migraine presumes that they share the underlying pathophysiology related to alterations in sodium and calcium ion channels and ion transporters (sodium-potassium transport; 8, 9). Moreover, drugs acting on voltage-gated sodium channels and γ-aminobutyric acid (GABA) receptors (e.g. valproate, topiramate, etc.) are used not only prevent migraine attacks, but are also used as anti-epileptic drugs (10, 11).

The prevalence of epileptic seizures in AD patients is strongly influenced by genetic factors (12)—epilepsy and seizures occur more often in patients with early-onset familial AD than those with sporadic AD (13). Further, convulsive seizures have been described in approximately 40% of familial AD patients with the PSEN1 p.Glu280Ala mutation (14), 30% with PSEN2 mutations (15) and 57% with amyloid precursor protein (APP p.Thr174Ile) duplications (16). Additionally, the p.Pro86Lys mutation in CALHM1 (rs11191692) is associated with both AD and temporal lobe epilepsy through its influence on calcium homeostasis (17).

Evidence that some comorbidities (e.g. AD and depression) also act as risk factors for developing seizures suggests they may share pathophysiological mechanisms (13, 18). Conversely, epileptic seizures (as well as neoplasms) have been reported to cause intellectual disabilities in patients with tuberous sclerosis (19, 20). Furthermore, certain anti-epileptic drugs (e.g. topiramate) are associated with higher incidence of cognitive problems (1, 21).

The several causal and associative relationships observed between epilepsy and other indications on the phenotypic level warrant further investigation for shared elements on the genetic and molecular levels. While inquiry on the genetic level often begins with genome-wide association studies to identify shared loci, identifying the appropriate data set(s) and linking intergenic single nucleotide polymorphisms (SNP) to their functional consequences across scales in complex disease is still a significant challenge (22). Even after identifying disease-associated genes, it is difficult to assess their individual contributions to the complex etiology of epilepsy and its related indications. Thus, capturing the different causal relationships between biological entities involved in the pathophysiology of a complex disease is an essential step to a better understanding of the processes that lead to the disease state.

Here, we present two methods to hypothesize shared mechanisms: first, a literature-based approach for quantifying gene overlap between epilepsy and its comorbidities as a proxy for a shared genetic basis of disease, or genetic pleiotropy; and second, a systematic approach using the NeuroMMSig mechanism enrichment server. Finally, we use these methods to propose an explanation for the observed therapeutic effects of a drug that has been studied in the contexts of epilepsy and AD. A schematic representation of the methodology, analysis and results is presented in Figure 1.

Materials and methods

Pre-processing of epidemiological studies

Epidemiological studies comparing the incidence of several comorbidities of epilepsy versus incidence in the general population were extracted from a recent review by Keezer *et al.* (23). The prevalence ratio, which describes the ratio of incidence of a given condition in epilepsy patients versus the general population, was calculated for each study.

Quantification of gene overlap

SCAIView v1.7.3, indexed from MEDLINE on 2016-07-14, (http://academia.scaiview.com/academia) was used to identify and quantify the overlap of genes co-occurring in the literature with epilepsy and its novel and well-known comorbidities reviewed by Keezer *et al.* Three comorbidities were excluded from assessment by literature-based methods due to their poor correspondence with MeSH terms (i.e. allergie) or their lack of specificity (i.e. heart disease and neoplasms) by occurring too frequently and covering too many genes to be insightful. Finally, additional epidemiological comorbidity studies with PD were curated and included due to its previously published relevance (24).

In order to later assess publication bias in literature-based methods, the total number of documents associated with each disease was reported by querying SCAIView

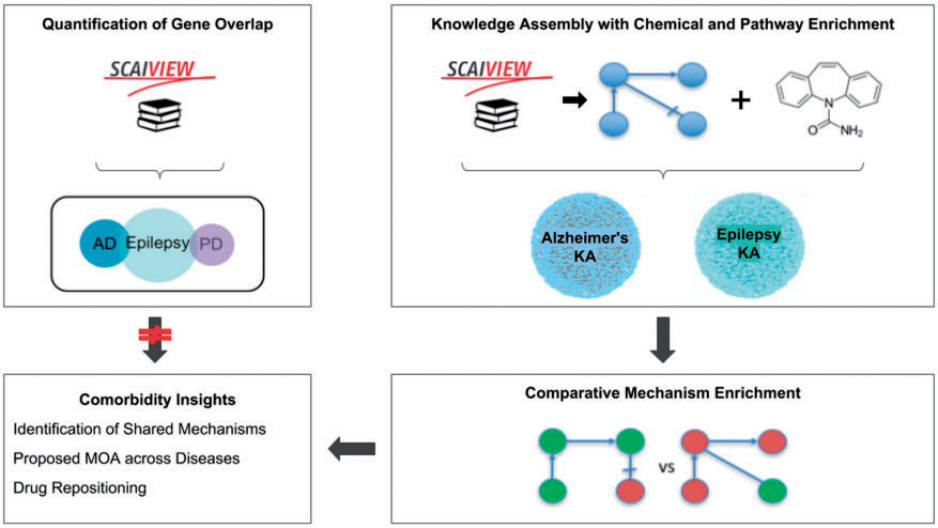


Figure 1. A graphical abstract of the methodology, analysis, and results presented in this work. The two upper boxes represent the methodology while the two lower boxes represent the analysis and results. The upper-left box outlines the quantification of gene overlap between epilepsy and its well-known comorbidities described in Keezer *et al.* (23; e.g. AD, PD, etc.) using literature based methods. The upper-right box outlines the assembly of knowledge from epilepsy literature with chemical and pathway enrichment as described in the 'Preparation for Mechanism Enrichment' and 'Relation Extraction' Sections. The lower-right box represents the comparative mechanism enrichment that was used to generate comorbidity insights (lower-left box) after literature-based methods proved insufficient.

with each disease's corresponding MeSH term. The total number of associated genes with each disease was counted by those with a positive relative entropy (i.e. occur more frequently in the results of a given SCAIView query than in the rest of the SCAIView indexed literature) in the context of the query as described by Younesi *et al.* (25). Gene sets for each comorbidity were then retrieved by constructing queries using their corresponding MeSH terms joined with the epilepsy MeSH term by the 'AND' operator. The associated genes for each comorbidity query were identified with the same method and the epilepsy pleiotropy rate was calculated as the percentage of the genes associated with the comorbidity query in the set of associated genes with epilepsy (Table 1).

For example, 226 genes were found to be associated with diabetes using the comorbidity query `[MeSH Disease:"Epilepsy"] AND [MeSH Disease:"Diabetes Mellitus"]`. Of these, 184 had positive relative entropies, which indicate that these genes occur more frequently in literature mentioning both diseases than in the rest of the SCAIView indexed literature. Finally, the epilepsy pleiotropy rate was calculated normalizing 184 by the total number of genes (2901) with a positive relative entropy found by querying for epilepsy `[MeSH Disease:"Epilepsy"]`, resulting in an epilepsy pleiotropy rate of 6.34%.

Relation extraction

While literature co-occurrence can generate initial hypotheses about genetic pleiotropy, it does not provide sufficient mechanistic insight to explain the clinical observations in epilepsy and its comorbidities. The increasing quantity of knowledge in the biomedical domain makes it difficult or impossible for researchers to be knowledgeable in any but incredibly specific topics (26). The task of manually generating pleiotropy hypotheses that explain overlap between the aetiological mechanisms of epilepsy and its comorbidities is daunting. In order to enable computer-aided automatic reasoning, the knowledge surrounding epilepsy and its comorbidities was systematically extracted from the literature using manual relation extraction and encoded in Biological Expression Language (BEL; 27).

First, a corpus was generated from the 192 245 documents related to epilepsy retrieved (Table 1) to further investigate the causal relations surrounding the identified genes. A second corpus was generated from the 2666 documents retrieved by querying SCAIView for 'epilepsy' and its sub-terms in the Epilepsy Ontology (28) occuring with the free text, 'comorbidity'. Manual relation extraction and encoding in BEL was then performed starting with a select subset of the two corpora based on their prioritization by SCAIView to generate the Epilepsy Knowledge Assembly. The knowledge assembly was further enriched

Table 1. Results of the Epilepsy comorbidity analysis using SCAIView

Disease (MeSH ID)	Associated documents	Disease associated genes	Comorbidity associated genes	Epilepsy pleiotropy rate (%)
Epilepsy (D004827)	192 245	2901	—	—
Stroke (D020521)	210 846	4533	633	17.78
AD (D000544)	109 495	4968	396	13.65
Migraine (D008881)	30 928	1230	306	10.54
PD (D010300)	79 103	3646	258	8.89
Hypertension (D006973)	391 190	5574	252	8.68
Dementia (D003704)	183 802	5833	220	7.58
Diabetes mellitus (D003920)	394 411	6661	184	6.34
Intestinal diseases (D007410)	629 691	9093	166	5.72
Thyroid diseases (D013959)	153 025	4366	133	4.58
Anxiety (D001007)	84 138	1782	124	4.27
Arthritis (D001168)	259 327	5367	122	4.2
Cataract (D002386)	52 150	2238	119	4.1
Asthma (D001249)	147 697	3761	86	2.96
Glaucoma (D005901)	56 679	2303	48	1.65
Depressive disorder, major (D003865)	15 706	1249	46	1.58
Urinary incontinence (D014549)	34 170	720	24	0.82
Peptic ulcer (D010437)	68 234	1445	21	0.72
Back pain (D001416)	48 516	1191	17	0.58
Pulmonary disease, chronic obstructive (D029424)	35 627	2244	15	0.51
Fibromyalgia (D005356)	9021	468	10	0.34
Emphysema (D004646)	25 511	1261	9	0.31
Bronchitis, chronic (D029481)	9085	580	2	0.06

Notes: The number of ***associated documents*** (column 2) retrieved from SCAIView for each disease is shown given a reference query using corresponding the MeSH term from column 1. The ***disease-associated genes*** (column 3) contain the number of genes relevant to the corpus retrieved from a disease-specific query. The ***comorbidity-associated genes*** (column 4) contain the number of genes relevant to the comorbidity query between the target disease and epilepsy. Lastly, the ***epilepsy pleiotropy rate*** (column 5) describes the ratio of the count of genes reported in column 4 with the total number of epilepsy-associated genes (2901).

with pharmacological knowledge surrounding 19 anti-epileptic drugs and their targets from the Pharmacogenetics and Pharmacogenomics Knowledge Base (PharmGKB; 29). Ultimately, the knowledge assembly comprised relations from 641 unique citations. Finally, the PyBEL framework (30) was used to parse and validate the syntax and semantics of the underlying BEL Script. A summary of the contents of the Epilepsy Knowledge Assembly is presented in Table 2.

Preparation for mechanism enrichment

The Epilepsy Knowledge Assembly was enriched with mechanistic annotations following the procedures outlined by Domingo-Fernández *et al.* (24) in order to integrate it into NeuroMMSig and enable multi-modal mechanism enrichment analyses with queries over genes, SNPs and neuroimaging features.

A taxonomy of epilepsy mechanisms was generated by combining the list of well-established epilepsy mechanisms from Staley (31) with concepts from the Pathway Terminology System (32) co-occurring in articles matching either [MeSH Disease: 'Epilepsy'] or entries in the Epilepsy Ontology indexed by SCAIView. The resulting 784 terms representing mechanisms were curated in order to normalize entities, remove irrelevant entries and group similar terms. Next, relations in the Epilepsy Knowledge Assembly were annotated with mechanisms based on whether their entities were involved in the mechanism as outlined by Domingo-Fernández *et al.* (24). During the annotation process, new mechanisms not yet included in the inventory were found in the literature; thus, the mechanism inventory was updated in parallel until concluding the annotation with a total of 32 annotated sub-graphs (Table 2). More details and examples about the mapping procedure can be found on the NeuroMMSig introduction page.

In the next section, NeuroMMSig is used to identify shared mechanisms between epilepsy and AD on the basis of the mechanism of action of multi-indication drugs.

Results and discussions

Investigation of comorbidities

Of the 2901 genes identified as associated with epilepsy by having a positive relative entropy score, Table 1 shows

Table 2. Statistics over the Epilepsy Knowledge Assembly generated by PyBEL, grouped by sub-graph

Sub-graph name	Biological entities	Relationships	Connected Components	Citations
Adaptive immune system sub-graph	12	12	4	5
Adenosine signaling sub-graph	76	154	3	15
Apoptosis signaling sub-graph	228	503	5	115
Brain-derived neurotrophic factor signaling sub-graph	75	142	1	29
Calcium dependent sub-graph	302	793	8	73
Chromatin organization sub-graph	8	10	2	2
Energy metabolic sub-graph	91	177	4	24
Estradiol metabolism	7	8	2	1
G-protein-mediated signaling	78	140	5	25
Gaba sub-graph	262	632	2	56
Glutamatergic sub-graph	121	246	5	32
Hormone signaling sub-graph	126	256	9	16
Inflammatory response sub-graph	42	49	4	21
Innate immune system sub-graph	41	63	8	18
Interleukin signaling sub-graph	46	111	3	17
Long term synaptic depression	64	129	2	21
Long term synaptic potentiation	125	252	2	46
Mapk-erk sub-graph	313	706	5	68
Metabolism	340	600	14	126
Mirna sub-graph	5	4	1	3
Mossy fiber sub-graph	38	66	2	14
Mtor signaling sub-graph	166	336	3	42
Neurotransmitter release sub-graph	552	1667	5	131
Notch signaling sub-graph	105	205	3	20
Protein kinase signaling sub-graph	377	850	6	87
Protein metabolism	129	185	8	44
Reelin signaling sub-graph	117	253	2	21
Regulation of actin cytoskeleton sub-graph	5	3	2	2
Serotonergic sub-graph	148	478	2	18
Thyroid hormone signaling sub-graph	106	228	2	9
Transport related sub-graph	49	60	7	25
Wnt signaling sub-graph	27	38	3	15
Total	*3478*	*12481*	*16*	*641*

Notes: The first column, ***biological entities***, corresponds to the number of genes, chemicals, proteins, biological process, etc. in each sub-graph. The second column, ***relationships*** (i.e. edges), corresponds to the number of connections between each sub-graphs' biological entities. The third column, ***connected components***, corresponds to the number of 'connected' groups of nodes within each sub-graph. The final column, ***citations***, corresponds to the total number of articles from which information was extracted to build each sub-graph. A more detailed summary is included in the Supplementary Material.

that nervous system conditions were among the highest literature-based gene overlap (10.54% of epilepsy genes co-occurred with migraine, 7.58% with dementia, 8.89% with PD and 13.65% with AD). While these conditions were also highly ranked in epidemiological studies by their prevalence ratios, literature-based gene overlap does not correlate with prevalence ratios across all of the conditions reported by Keezer *et al.* (Supplementary Figure S2) and therefore is the best tool for gaining insight into the comorbidities of epilepsy.

While literature-based methods may be a poor proxy for genetic pleiotropy and are generally insufficient for unraveling the shared pathophysiology in epilepsy and studied comorbidities, systematic approaches for identifying and evaluating shared mechanisms may better explain the aggregate effects of their interactions that cannot be captured by simple approaches like literature co-occurrence.

Mechanism enrichment

AD was chosen as a putative comorbidity of epilepsy for further investigation not only because it had the highest literature-based gene overlap with epilepsy (Table 1), but additionally because of the prior existence of the AD Knowledge Assembly (33) and its inclusion in NeuroMMSig.

Of the most frequently mentioned drugs in epilepsy literature (Supplementary Figure S1), we identified carbamazepine as having both notable target representation in the AD Knowledge Assembly along with multiple references indicating its positive effects on memory and treatment of elderly patients with seizures (34) as well as positive effects in the treatment of AD patients (34, 35).

In the following sections, NeuroMMSig is first used to investigate the mechanisms enriched by the targets of carbamazepine in the context of epilepsy. After, a comparative enrichment is made to identify possible overlapping mechanisms with AD in order to explore the therapeutic effects of the drug in both disease contexts.

Epilepsy mechanism enrichment

While carbamazepine has been observed to act through inhibition of sodium and calcium voltage-gated channels as well as activation of the GABA receptors (36, 37), its mechanism of action is still not fully understood (38). In order to better understand the impact of the drug, the gene set of all of its known targets (Supplementary Text S1), was queried on the NeuroMMSig mechanism enrichment server against the Epilepsy Knowledge Assembly.

Because the NeuroMMSig mechanism enrichment algorithm only provides relative scores, the top 10th percentile of results were used. Of the 12 networks with at least one mapped gene, those with an enrichment score in the top 10th percentile (above 0.696) were the adenosine signaling sub-graph and the GABA sub-graph (Supplementary Table S1). While the increase of the purine nucleoside, adenosine, has been associated with the incidence of seizures, its mechanistic connection is still unknown (39). However, recent research has identified several promising targets that regulate and balance adenosine levels such as adenosine kinase (ADK) and its receptors (ADORA family; 39, 40). Similarly to adenosine, the inhibitory neurotransmitter response induced by GABA is responsible for balancing many excitatory signals occurring in the brain. Studies that investigated reduction and abnormalities in GABA-inhibitory processes lead to the development of GABA agonists (e.g. vigabatrin and tiagabine) that act as anti-convulsants in epilepsy patients (41).

After assessing the plausibility of these sub-graphs' involvement in the aetiology of epilepsy, the union of the networks was used to further investigate the relation between the downstream effects of carbamazepine and its therapeutic effect on epilepsy patients. Due to the its size, it was necessary to filter and query the network by finding the shortest paths between the drug's targets and different biological processes of interest in the context of epilepsy then combining them to form a new graph to support quickly identifying candidate pathways. Finally, the common upstream controllers between each pathway were included to provide further context to their overlap. Combining, reasoning over, and manually interpreting the generated paths lead to the simplified network depicted in Figure 2 representing the downstream effects of carbamazepine.

The figure demonstrates the synergistic effects of the activation of the GABA receptor family and the inhibition of the sodium voltage-gated channels to potentiate GABA-mediated inhibition, and therefore, decrease the risk of developing seizures. Furthermore, carbamazepine causes an increase in the production of adenosine, whose receptors, ADORA1 and ADORA3, in a positive feedback loop with the production of adenosine which ultimately leads to a reduction in glutamatergic excitatory signals. Synaptic plasticity in response to the downstream effects of these signals may contribute to drug resistance, and ultimately, seizure recurrence.

Comparative mechanism enrichment

NeuroMMSig was queried with the same gene set in the context of AD in order to perform a comparative investigation of shared mechanistic perturbations with epilepsy. Because the mechanism of action of carbamazepine is poorly represented in the literature (Supplementary Table S2) and its known targets are less implicated in AD, it was unsurprising that fewer sub-graphs were enriched in the context of AD.

The most significant, the GABA sub-graph, which describes the upstream controllers of the GABA receptor and its downstream effectors, highly overlapped with the GABA sub-graph in the context of the epilepsy–it contains key relations that may explain the efficacy of carbamazepine in both conditions. Studies in AD models have shown a negative correlation between the abundance of amyloid beta 42 and the expression of the transcription factor for the GABA receptor family EGR1 (42, 43). This correlation could be caused by an unknown controller in the aetiology of AD; in which state, a patient would have decreased expression of EGR1 and therefore fewer GABA receptors and less ability to inhibit the excitatory signals that lead to seizures. While the link between epilepsy and GABAergic neurotransmission has already been exploited by anti-epileptic drugs, its role in AD is not yet clearly understood (Figure 2). However, several recent publications have rationalized targeting GABAergic neurotransmission for treatment of AD (44, 45).

Tangentially, EGR1 upregulates the expression of PSEN2, a member of the catalytic sub-unit of the γ-secretase complex that regulates APP cleavage (46). Mutations

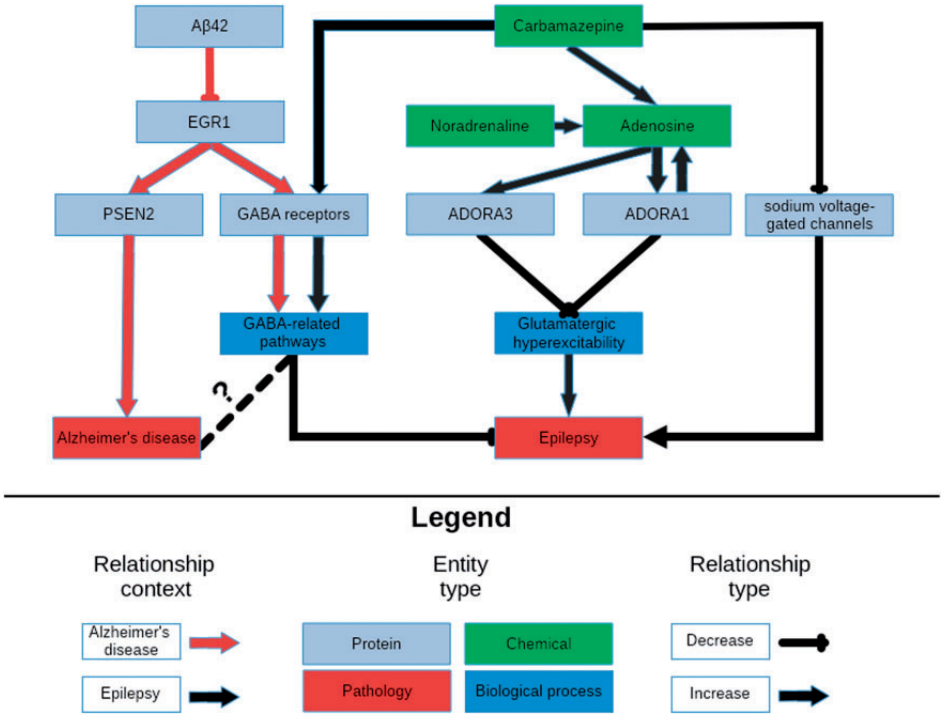


Figure 2. A schematic representation of the knowledge surrounding carbamazepine retrieved by querying its targets with NeuroMMSig. The relevant portions of the most significantly enriched graphs in the context of epilepsy and AD, the adenosine signaling and the GABA sub-graphs, were merged and displayed in order to highlight a potential explanatory mechanism for the therapeutic effects of carbamazepine (Supplementary Text S2). It is rendered with a hierarchical layout to mirror the flow from molecular entities to proteins, biological processes and pathologies.

in PSEN2 that have been both linked to amyloid beta accumulation (47) and seizures in AD patients (48) provide further evidence for the existence for a shared mechanism through which carbamazepine acts in the contexts of AD and epilepsy (Figure 2).

Noradrenaline is a known anti-convulsant (49) that is often lacking in patients in the early stages of AD due to an observed loss of noradrenergic neurons (50). Because carbamazepine has been observed to activate noradrenergic neurons (51), its anti-convulsant activity may be due to it indirectly increasing noradrenaline levels. Finally, noradrenaline potentiates the previously mentioned adenosine pathways (52).

While mechanism enrichment provides several insights, a cursory search of PubMed of 'Carbamazepine'[nm] AND 'Alzheimer disease'[MeSH Terms] suggested publications (53) that implicate autophagy in the therapeutic action of Carbamazepine. Further investigation showed Carbamazepine does not have a significant effect on expression of proteins in the mTOR-pathway (53) and that it is likely increasing autophagic flux through an mTOR-independent pathway. Mechanism enrichment analysis was unable to prioritize autophagy pathways due to some of the shortcomings of knowledge-based methods. For example, the AD NeuroMMSig sub-graph corresponding to autophagy pathways did not contain any of the targets of Carbamazepine listed by PharmGKB. This could be due to the choice of boundaries in sub-graph definition, or also due to the lack of annotation of autophagy-related targets in PharmGKB. Autophagy has been implicated in epilepsy, but the literature has not yet succinctly described the connection from the therapeutic to its target, pathway, and finally the pathology. Finally, because knowledge assemblies are inherently incomplete, this shows the complementary nature of the two approaches.

While the exact mechanism of action of carbamazepine remains elusive, this proposed mechanism enrichment approach was able to identify multiple pathways through which it could be acting in both the AD, epilepsy and shared context. Looking forward, this approach can be applied across a wide variety of chemical matter in the neurodegenerative disease space, as well as in other domains for which appropriately annotated knowledge assemblies exist, in order to support identification of drugs' mechanisms of actions, drug repositioning opportunities and the development of new lead compounds.

Conclusion

Our findings indicate that literature-based methods as a proxy for genetic pleiotropy generally do not correlate with the results from epidemiological studies of epilepsy and its comorbidities. Furthermore, strictly gene-centric methods lack the ability to elucidate mechanistic insight that a knowledge assembly can support.

After formalizing a representative sample of the knowledge surrounding epilepsy, its risk factors, its comorbidities and anti-epileptic drugs, we annotated mechanistic subgraphs to include in the NeuroMMSig mechanism enrichment server. Finally, an enrichment approach focusing on the targets of carbamazepine proposed the several downstream effectors (including the GABA receptor, GABAergic pathways, etc.) that could explain its therapeutic effects in both the contexts of epilepsy and AD.

Future work will include applying this procedure to a wider variety of drugs and chemical matter across different diseases. Finally, we have made the Epilepsy Knowledge Assembly publicly available through NeuroMMSig (http://neurommsig.scai.fraunhofer.de) to facilitate further systems biology and chemoinformatics investigations of epilepsy.

Supplementary data

Supplementary data are available at *Database* Online.

Acknowledgement

The authors would like to thank our colleagues at Fraunhofer SCAI for their critical review of this manuscript.

Funding

This work was supported by the European Union/European Federation of Pharmaceutical Industries and Associations (EFPIA) Innovative Medicines Initiative Joint Undertaking under AETIONOMY [Joint Technology Initiatives 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

1. Baillieux,H., Verslegers,W., Paquier,P. *et al*. (2008) Cerebellar cognitive a ective syndrome associated with topiramate. *Clin. Neurol. Neurosurg*., **110**, 496–499.
2. Fisher,R.S., Acevedo,C., Arzimanoglou,A. *et al*. (2014) ILAE o cial report: a practical clinical de nition of epilepsy. *Epilepsia*, **55**, 475–482.
3. Téllez-Zenteno,J.F., Matijevic,S. and Wiebe,S. (2005) Somatic comorbidity of epilepsy in the general population in Canada. *Epilepsia*, **46**, 1955–1962.
4. Sander,J.W. (2013) Comorbidity and premature mortality in epilepsy. *Lancet*, **382**, 1618–1619.
5. Selassie,A.W., Wilson,D.A., Martz,G.U. *et al*. (2014) Epilepsy beyond seizure: a population-based study of comorbidities. *Epilepsy Res*., **108**, 305–315.
6. Seidenberg,M., Pulsipher,D.T. and Hermann,B. (2009) Association of epilepsy and comorbid conditions. *Future Neurol*., **4**, 663–668.
7. Toldo,I., Perissinotto,E., Menegazzo,F. *et al*. (2010) Comorbidity between headache and epilepsy in a pediatric headache center. *J. Headache Pain*, **11**, 235–240.
8. Winawer,M.R. and Connors,R. (2013) Evidence for a shared genetic susceptibility to migraine and epilepsy. *Epilepsia*, **54**, 288–295.
9. Rogawski,M.A. (2012) Migraine and epilepsy—shared mechanisms within the family of episodic disorders. In: Noebels,J.L., Avoli,M., Rogawski,M.A. *et al*. (eds). *Jasper's Basic Mechanisms of Epilepsies*, 4th edn. National Center for Biotechnology Information, Bethesda, MD, pp. 930–944.
10. Storey,J.R., Calder,C.S., Hart,D.E. *et al*. (2001) Topiramate in migraine prevention: a double-blind, placebo-controlled study. *Headache*, **41**, 968–975.
11. Rothrock,J.F. (2012) Topiramate for migraine prevention: an update. *Headache*, **52**, 859–860.
12. Miranda,D.D.C. and Brucki,S.M.D. (2014) Epilepsy in patients with Alzheimer's disease: a systematic review. *Dement Neuropsychol*., **8**, 66–71.
13. Amatniek,J.C., Hauser,W.A., DelCastillo-Castaneda,C. *et al*. (2006) Incidence and predictors of seizures in patients with Alzheimer's disease. *Epilepsia*, **47**, 867–872.
14. Larner,A.J. and Doran,M. (2006) Clinical phenotypic heterogeneity of Alzheimer's disease associated with mutations of the presenilin–1 gene. *J. Neurol*., **253**, 139–158.
15. Jayadev,S., Leverenz,J.B., Steinbart,E. *et al*. (2010) Alzheimer's disease phenotypes and genotypes associated with mutations in presenilin 2. *Brain*, **133**, 1143–1154.
16. Cabrejo,L., Guyant-Maréchal,L., Laquerrière,A. *et al*. (2006) Phenotype associated with APP duplication in ve families. *Brain*, **129**, 2966–2976.
17. Lv,R.J., He,J.S., Fu,Y.H. *et al*. (2011) A polymorphism in CALHM1 is associated with temporal lobe epilepsy. *Epilepsy Behav*., **20**, 681–685.
18. Palop,J.J. and Mucke,L. (2009) Epilepsy and cognitive impairments in Alzheimer disease. *Arch. Neurol*., **66**, 435–440.
19. Jansen,F.E., Vincken,K.L., Algra,A. *et al*. (2008) Cognitive impairment in tuberous sclerosis complex is a multifactorial condition. *Neurology*, **70**, 916–923.
20. Osborne,J.P., Lux,A.L., Edwards,S.W. *et al*. (2010) The underlying etiology of infantile spasms (West syndrome): information from the United Kingdom Infantile Spasms Study (UKISS) on contemporary causes and their classification. *Epilepsia*, **51**, 2168–2174.
21. Thompson,P.J., Baxendale,S.A., Duncan,J.S. *et al*. (2000) E ects of topiramate on cognitive function. *J. Neurol. Neurosurg. Psychiatry*, **69**, 636–641.

22. Lee,S.H., Yang,J., Goddard,M.E. *et al*. (2012) Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics*, **28**, 2540–2542.
23. Keezer,M.R., Sisodiya,S.M. and Sander,J.W. (2016) Comorbidities of epilepsy: current concepts and future perspectives. *Lancet Neurol*., **15**, 106–115.
24. Domingo-Fernández,D., Kodamullil,A.T., Iyappan,A. *et al*. (2017) Multimodal mechanistic signatures for neurodegenerative diseases (NeuroMMSig): a web server for mechanism enrichment. *Bioinformatics*, **33**, 3679–3681.
25. Younesi,E., Toldo,L., Müller,B. *et al*. (2012) Mining biomarker information in biomedical literature. *BMC Med. Inf. Decision Making*, **12**, 148.
26. Bellazzi,R. (2014) Big data and biomedical informatics: a challenging opportunity. *Yearbook Med. Inf*., **9**, 8.
27. Slater,T. (2014) Recent advances in modeling languages for pathway maps and computable biological networks. *Drug Discov. Today*, **19**, 193–198.
28. Almeida,P., Gomes,P., Sales,F. *et al*. (2010) Ontology and knowledge management system on epilepsy and epileptic seizures. *Eprint arXiv*, 1012.1638.
29. Whirl-Carrillo,M., McDonagh,E.M., Hebert,J.M. *et al*. (2012) Pharmacogenomics knowledge for personalized medicine. *Clin. Pharmacol. Therapeutics*, **92**, 414–417.
30. Hoyt,C.T., Konotopez,A. and Ebeling,C. (2017) PyBEL: a computational framework for Biological Expression Language. *Bioinformatics*, btx660.
31. Staley,K. (2015) Molecular mechanisms of epilepsy. *Nat. Neurosci*., **18**, 367–372.
32. Iyappan,A., Gündel,M., Shahid,M. *et al*. (2016) Towards a pathway inventory of the human brain for modeling disease mechanisms underlying neurodegeneration. *J. Alzheimer's Dis*., **52**, 1343–1360.
33. Kodamullil,T.A., Younesi,E., Naz,M. *et al*. (2015) Computable cause-and-e ect models of healthy and Alzheimer's disease states and their mechanistic di erential analysis. *Alzheimer's & Dementia: J. Alzheimer's Assoc*., **11**, 1329–1339.
34. Li,L., Zhang,S., Zhang,X. *et al*. (2013) Autophagy enhancer carbamazepine alleviates memory deficits and cerebral amyloid-β pathology in a mouse model of Alzheimer's disease. *Curr. Alzheimer Res*., **10**, 433–441.
35. Tariot,P.N., Erb,R., Podgorski,C.A. *et al*. (1998) E cacy and tolerability of carbamazepine for agitation and aggression in dementia. *Am. J. Psychiatry*, **155**, 54–61.
36. Granger,P., Biton,B., Faure,C. *et al*. (1995) Modulation of the gamma-aminobutyric acid type A receptor by the antiepileptic drugs carbamazepine and phenytoin. *Mol. Pharmacol*., **47**, 1189–1196.
37. Liu,L., Zheng,T., Morris,M.J. *et al*. (2006) The mechanism of carbamazepine aggravation of absence seizures. *J. Pharmacol. Exper. Therapeutics*, **319**, 790–798.
38. Ambrósio,A.F., Soares-da-Silva,P., Carvalho,C.M. *et al*. (2002) Mechanisms of action of carbamazepine and its derivatives, oxcarbazepine, BIA 2-093, and BIA 2-024. *Neurochem. Res*., **27**, 121–130.
39. Boison,D. (2013) Adenosine and seizure termination: endogenous mechanisms. *Epilepsy Curr*., **13**, 35–37.
40. Masino,S.A., Kawamura,M. Jr. and Ruskin,D.N. (2014) Adenosine receptors and epilepsy: current evidence and future potential. *Int. Rev. Neurobiol*., **119**, 233–255.
41. Treiman,D.M. (2001) GABAergic mechanisms in epilepsy. *Epilepsia*, **42 Suppl 3**, 8–12.
42. Murphy,M.P. and LeVine,H. III. (2010) Alzheimer's disease and the amyloid-beta peptide. *J. Alzheimers Dis*., **19**, 311–323.
43. Mo,J., Kim,C.H., Lee,D. *et al*. (2015) Early growth response 1 (Egr- 1) directly regulates GABAA receptor? 2,? 4, and? subunits in the hippocampus. *J. Neurochem*., **133**, 489–500.
44. Solas,M., Puerta,E. and Ramirez,M.J. (2015) Treatment options in Alzheimer ś disease: the GABA story. *Curr. Pharm. Des*., **21**, 4960–4971.
45. Li,Y., Sun,H., Chen,Z. *et al*. (2016) Implications of GABAergic neurotransmission in Alzheimer's disease. *Front. Aging Neurosci*., **8**, 31.
46. Renbaum,P., Beeri,R., Gabai,E. *et al*. (2003) Egr-1 upregulates the Alzheimer's disease presenilin-2 gene in neuronal cells. *Gene*, **318**, 113–124.
47. Steiner,H., Capell,A., Leimer,U. *et al*. (1999) Genes and mechanisms involved in β-amyloid generation and Alzheimer's disease. *Eur. Arch. Psychiatry Clin. Neurosci*., **249**, 266–270.
48. Noebels,J. (2011) A perfect storm: converging paths of epilepsy and Alzheimer's dementia intersect in the hippocampal formation. *Epilepsia*, **52**, 39–46.
49. Giorgi,F.S., Pizzanelli,C., Biagioni,F. *et al*. (2004) The role of norepinephrine in epilepsy: from the bench to the bedside. *Neurosci. Biobehav. Rev*., **28**, 507–524.
50. Braak,H., Thal,D.R., Ghebremedhin,E. *et al*. (2011) Stages of the pathologic process in Alzheimer disease: age categories from 1 to 100 years. *J. Neuropathol. Exper. Neurol*., **70**, 960–969.
51. Olpe,H.R. and Jones,R.S. (1983) The action of anticonvulsant drugs on the ring of locus coeruleus neurons: selective, activating e ect of carbamazepine. *Eur. J. Pharmacol*., **91**, 107–110.
52. Pearson,T. and Frenguelli,B.G. (2004) Adrenoceptor subtype-speci c acceleration of the hypoxic depression of excitatory synaptic transmission in area CA1 of the rat hippocampus. *Eur. J. Neurosci*., **20**, 1555–1565.
53. Zhang,L., Wang,L., Wang,R. *et al*. (2017) Evaluating the effectiveness of GTM-1, rapamycin, and carbamazepine on autophagy and Alzheimer disease. *Med. Sci. Monitor: Int. Med. J. Exper. Clin. Res*., **23**, 801–808.

Conclusions

9 Conclusion and Outlook