# Capturing scientific knowledge in computable form

Jeffrey V. Wong[1]*, Max Franz[1]*, Metin Can Siper[2], Dylan Fong[1], Funda Durupinar[3], Christian Dallago[5,6,7], Augustin Luna[4,5,8], John Giorgi[1], Igor Rodchenkov[1], Özgün Babur[3], John A. Bachman[9], Benjamin M. Gyori[9], Emek Demir[2,*], Gary D. Bader[1,10,11]* and Chris Sander[4,5,8]*

[1] The Donnelly Centre, University of Toronto, Toronto, Ontario, M5S 3E1, Canada

[2] Computational Biology Program, Oregon Health and Science University, Portland, OR 97239, USA

[3] Computer Science Department, University of Massachusetts Boston, 100 Morrissey Boulevard Boston, MA 02125

[4] Department of Data Sciences, Dana-Farber Cancer Institute, Boston, MA, 02215, USA

[5] Department of Cell Biology, Harvard Medical School, Boston, MA, 02215, USA

[6] Department of Systems Biology, Harvard Medical School, Boston, MA, 02215, USA

[7] Department of Informatics, Technische Universität München, 85748 Garching, Germany

[8] Broad Institute of MIT and Harvard, Boston, MA, 02142, USA

[9] Laboratory of Systems Pharmacology, Harvard Medical School, Boston, MA, 02115, USA

[10] Department of Computer Science, University of Toronto, Ontario, M5S 2E4, Canada

[11] Department of Molecular Genetics, University of Toronto, Ontario, M5S 1A8, Canada

*Correspondence to info@biofactoid.org reaches the principal authors.

**Availability**: Biofactoid server at https://biofactoid.org

**Keywords**: pathway analysis; curation tool; knowledge base

## ABSTRACT

Technological advances in computing provide major opportunities to accelerate scientific discovery. The wide availability of structured knowledge would allow us to take full advantage of these by enabling efficient human-computer interaction. Traditionally, biological knowledge is captured in publications and knowledge bases, however, the information in articles is not directly accessible to computers, and knowledge bases are constrained by finite resources available for manual curation. To accelerate knowledge capture and communication and to keep pace with the rapid growth of scientific reports, we developed the Biofactoid (biofactoid.org) software suite, which crowdsources structured knowledge in articles from authors. Biofactoid is a web-based system that lets scientists draw a network of interactions between genes, their products, and chemical compounds and employs smart-automation to translate user input into a structured language using the expressive power of a formal ontology. The resulting data is shared via public information resources, enabling author-curated knowledge to be appreciated in the context of all existing computable knowledge. Authors of recently published papers across a range of journals have already contributed their pathway information, much of which is novel and extends existing pathway databases into new biological areas. We envision the adoption of Biofactoid for crowdsourced curation by scientists and publishers as part of an ecosystem of tools that accelerate scientific communication and discovery.

## INTRODUCTION

Biological pathways organize sets of molecular interactions and reactions that underlie cellular processes and are used for experimental design, interpretation of genomics data (Khatri et al., 2012), understanding of disease mechanisms (Chinen et al., 2016; Santos et al., 2014), and identification of therapeutic targets (Mack et al., 2014). To manage, visualize and interpret the large amount of available pathway information, researchers require computational tools such as pathway information resources, structured data representation standards, and analysis software (Demir et al., 2010; Franz et al., 2016; Jassal et al., 2020; Pratt et al., 2015; Rodchenkov et al., 2020; Shannon et al., 2003). These computational tools are capable of accessing increasing amounts of pathway and interaction data (Bader et al., 2006) collected through centralized (Gene Ontology Consortium, 2015; Jassal et al., 2020) or crowdsourced curation efforts (Slenter et al., 2018). Nevertheless, these existing efforts are not able to achieve wide coverage of the rapidly growing scientific corpus (over 1.3 million new PubMed articles/year) (Bornmann and Mutz, 2015; Cordero et al., 2016), arguing for the development of more scalable and sustainable approaches (Attwood et al., 2015; Imker, 2018).

For reasons of efficiency and accuracy, structured knowledge of biomedical discoveries could be provided directly by the original authors of research reports, rather than *post-hoc*, through the curation efforts of knowledge base teams. By analogy, molecular 3D structures submitted directly by authors to the Protein DataBank (PDB) (Berman et al., 2000) have become a key resource among structural biologists, and similar community practice led to direct submission of DNA sequence and transcriptomics information to public databases. Indeed, the importance and value of having such research outputs available in a community resource are underscored by the fact that data deposition is often a requirement of publication and funding. In contrast, there are few efforts and little technology to support

direct submission by authors of biological pathway information and related knowledge in computable form.

Here we introduce Biofactoid (biofactoid.org), a web-based software system that empowers authors to capture and share structured human- and machine-readable summaries of molecular-level interactions described in their publications. Without such a curation support tool, the onus would be on authors to handle a series of complex tasks involved in converting their knowledge to a computational form and depositing it in a suitable knowledge base. To overcome this and other significant barriers to computable knowledge acquisition and sharing, we developed Biofactoid to ease pathway curation and to rapidly generate expressive, structured representations with minimal user training. Structured knowledge newly acquired in this way becomes part of the global pool of pathway knowledge and can be shared in resources such as Pathway Commons (Rodchenkov et al., 2020), Network Data Exchange (NDEx) (Pratt et al., 2015), and STRING (Szklarczyk et al., 2017), to enhance information discovery and analysis. Authors can use Biofactoid to share structured information from research articles both as part of the publication process and outside of it. The development of Biofactoid and related computational tools helps support human-computer communication and inference algorithms in an ecosystem in which scientific reasoning is increasingly assisted by broad and deep knowledge computation.

## RESULTS

### Data sharing workflow

Biofactoid enables molecular-level detail of biological processes reported in articles to be shared in a structured format accessible to humans and computers. Interactions (including binding, post-translational modification, and transcription/translation) involving molecules of various types (proteins, nucleic acids, genes, or chemicals, e.g., metabolites and drug compounds) can be represented. Users begin by entering the article title, or identifier (e.g. PubMed identifier or Digital Object Identifier). Article metadata including authors, abstract and journal issue is automatically retrieved. Next, users draw a network of biological entities and interactions using the Biofactoid curation tool, which has an easy-to-use interface similar to graphical illustration software (e.g. Microsoft Powerpoint, Adobe Illustrator) which will be familiar to users, but with the added ability to generate structured data from the author-drawn biological pathway. The major features of the Biofactoid curation tool are as follows:

1. **Molecular entities.** Genes, gene products and chemicals are created in the network using an “Add a gene or chemical” tool, which creates a node (circle) that users label (Figure 1A). Biofactoid automatically matches the label with a record in an external database: NCBI Gene for genes and their products and ChEBI for small molecules (Brown et al., 2015; Hastings et al., 2016). This match represents the top hit of a search based on the similarity between a user’s label and a database record’s list of common names and synonyms; for genes, organisms are given priority based upon the organism of genes previously added to the network. Users can update the match by selecting another organism (e.g. human or mouse p53) or update the automatically inferred gene product type (i.e. RNA or protein). Alternatively, users may assign an alternate database entry from a list of search results. The system supports human and select model organisms (M. musculus, R.

norvegicus, S. cerevisiae, D. melanogaster, E. coli, C. elegans, D. rerio, and A. thaliana). The system is fast (response times of 100 milliseconds or less) and can be easily extended, for instance, to support SARS-CoV and SARS-CoV-2, which we added in response to the burst of publications related to the COVID-19 global pandemic (Ostaszewski et al., 2020). The matching process is also accurate, as measured using tests that use entity names from research articles as queries and assessing the quality of the search result. In over 90% of the cases, the correct result was first among search hits and was among the top 10 in over 97% of cases. Thus, Biofactoid can offer an accurate database identifier match using only the author-provided entity name. This represents a major advance in usability compared to traditional gene and chemical name querying systems that only work with exact name string matching and are often slow or unreliable. In fact, we were forced to build our own system after trying to use major existing systems that turned out to have high failure rates for our use case.

2. **Interactions.** A “draw interaction” tool lets users link two nodes by clicking one and dragging to the other (Figure 1B). Edges can have type and direction: a pointed arrow indicates stimulation or activation; a ‘T-bar’ arrowhead indicates inhibition or repression; and an undirected line indicates any other interaction. The mechanism can be refined by selecting from a list that includes binding, transcription/translation, or common forms of post-translational modifications, with additional interaction types to be added in future releases.

3. **Complexes.** Molecular complexes can be created by dragging genes or chemicals close to each other, resulting in a box that encloses them, which can also be labeled (Figure 1B).

4. **Automatic saving and co-editing.** Pathway representations in Biofactoid are automatically saved as changes are made and a live-sync capability enables multiple authors to collaboratively edit the same pathway, analogous to Google Docs.

Once complete, the pathway data is validated by pressing a ‘Submit’ button. At this point, the user may address any potential quality issues flagged by the system (e.g. unlabelled nodes, empty document) before confirming their submission.

## Data sharing and exploration

When research findings are shared through Biofactoid in a structured and computable manner, curated knowledge is automatically connected to, and becomes part of, a collective pool of computable knowledge that the community can access in different ways. Visitors to the Biofactoid website (biofactoid.org) may browse recently added articles, and a graphical abstract of each new submission is posted to Twitter (twitter.com/biofactoid). Each entry is automatically linked to its associated authors, article information and structured data and presented in an interactive Biofactoid Explorer web app (Figure 2A). Users can select any entity in the pathway diagram to see more information about genes (via NCBI Gene database links), proteins (via UniProt database links (UniProt Consortium, 2019)) and chemicals (via ChEBI database links). To support attribution, each author of an article is automatically linked to their profile in the Open Researcher and Contributor ID database (ORCID; orcid.org). Finally, a set of ‘Recommended articles’ are generated to help the user explore similar knowledge in the literature. To generate these recommendations, we first retrieve an article’s references, articles it is cited by and “Similar articles" from PubMed using the NCBI Entrez Programming Utilities service (Wheeler et al., 2006). These are combined

with articles that mention or provide evidence for interactions involving any of the molecular entities in the Biofactoid entry, which are accessed from curated pathway and interaction databases as well as resources that aggregate interaction information directly from the biomedical literature using natural language processing (NLP) tools (e.g. REACH, INDRA) (Gyori et al., 2017; Valenzuela-Escárcega et al., 2018). Recommended articles are ranked by determining how similar their titles and abstracts are to that of the Biofactoid article, using a deep-learning-based method we developed (Giorgi et al., 2020). The list of recommendations is context-sensitive in that articles are shown relevant to the part of the pathway selected by a user.

Beyond the 'Explorer', Biofactoid data represented in the formal BioPAX data exchange language is integrated with external pathway and interaction knowledge using technologies for BioPAX processing and analysis, including Paxtools (Demir et al., 2013) and cPath2 (Cerami et al., 2006). This enables author-contributed information to be more easily searched, visualized and analyzed across different data sources. For example, researchers can ask the question: "*Which known pathways are related to an interaction in Biofactoid?*" As an example answer to this question, a search of Pathway Commons shows how two interactions made computable by authors using Biofactoid (i.e. "*SENP1 activates SIRT3"* (T. Wang et al., 2019); "*SIRT5 activates UCP1*" (G. Wang et al., 2019)) unify previously disconnected pathways for mitochondrial biogenesis from the Reactome Pathway Database (Jassal et al., 2020) (Figure 2B). Researchers can also ask: "*How are two genes related?*" As another query example, a search of Pathway Commons shows how an interaction contributed by an author using Biofactoid ("*SENP1 activates SIRT3"* (T. Wang et al., 2019) ) provides a new, more direct regulatory pathway linking the mitochondrial proteins SENP1 and SOD2. Thus, Biofactoid directly contributes to building a more complete and comprehensive collection of biological pathways.

## Pilot study involving authors and journals

We tested the Biofactoid software over many iterations of testing involving authors and journal editors, extensively updating the software each time based on user feedback. This process supported two major goals: to improve the user experience of the software for authors, who are typically unfamiliar with curation and structured data concepts, and to develop a model for integrating Biofactoid into the publication process with journals. Once Biofactoid software achieved a sufficient level of sophistication and completeness, satisfying many of these initial users, we engaged journal editors to determine the feasibility of using Biofactoid to capture information from authors. Our pilot study consisted of three phases (Figure 3). Phase I introduced Biofactoid to journal editors via an "author-simulation", which began with a mock email invitation asking the editor to use Biofactoid to curate a selected article, and ending when a pathway from that article was input into the system by the editor using the Biofactoid web-based user interface. Phase II involved sending email invitations to a small number (N=15) of authors from a single journal, with multiple successful responses, proving that authors can use the system without any direct support from the Biofactoid team or journal. Phase III measured engagement rate by emailing 260 authors who published selected articles across 16 journals. Roughly 8.5% of these authors successfully shared their research in Biofactoid, proving that many authors can and will use Biofactoid (Figure 2A). We also found that 10% of articles across selected journals are suitable for inclusion in Biofactoid based on the current set of supported concepts (Table S1). In addition to proving

that Biofactoid can be independently used by authors to capture pathway knowledge, these results indicate a strong willingness among authors in the research community to use the system. These also demonstrate that the information captured by authors may not be present in any existing pathway database and helps connect previously unconnected entities and pathways in these databases (Figure 2B).

## DISCUSSION

Biofactoid focuses on the capture of published pathway information in computational form to augment discovery, attribution, and communication of scientific knowledge. In developing our strategy, we have considered how technology can best be used to aid authors. The result is a generic approach that can be extended to capture other types of biological knowledge, at the source of knowledge creation, in computable form. To be successful, Biofactoid development has focused on providing key elements of efficiency, incentives, and better technology for human-computer interaction. The computable knowledge capture model proposed here includes formal knowledge representation using ontologies, easy to use curation support software that links molecules to corresponding database identifiers (normalization), submission to widely available knowledge resources, a connection with the publication process and author attribution. This model expands on prior work defining digital abstracts and building crowdsourcing efforts for pathway data and software (Bharadwaj et al., 2017; Ceol et al., 2008; Gerstein et al., 2007; Liechti et al., 2017; Pratt et al., 2015; Slenter et al., 2018; Todorov et al., 2019; Waagmeester et al., 2020). Future development will include expanding the Biofactoid process to capture increasing amounts of literature-described pathway information such as additional relationship types, context and direct vs. indirect interactions, prioritized by frequency of occurrence in publications. In principle, Biofactoid technology can be extended to support any network of concepts and their relationships (i.e. knowledge graph) collaboratively built by a community of knowledge generators.

A convenient time to capture knowledge is during publication when authors - the primary source of knowledge - are most aware of the details of their report and when they are typically asked to deposit other types of data (e.g. sequencing or expression data). Thus, Biofactoid can integrate with the publication process, but also can be used in crowdsourcing efforts, which have been especially useful in curating knowledge about SARS-CoV-2 (Ostaszewski et al., 2020). As a demonstration, our pilot study engaged publishers to establish requirements for using Biofactoid within their publication pipeline, ideally as a condition for acceptance, and to drive greater awareness of the system among the wider readership. The pilot study also revealed that directly inviting authors of suitable articles to contribute, even after publication, is a viable engagement strategy. However, the time-consuming and labour-intensive nature of manually screening articles argues for the development of systems (e.g. using NLP) to automate article triage. If successful, this approach could be used to regularly identify relevant articles from the pool of all new entries indexed by PubMed. Another approach to reach users is to notify those whose research articles are referenced by information curated in Biofactoid. For this purpose, we have developed a system that automatically notifies authors when their research is linked to papers curated in Biofactoid (e.g. by citations or because of related content) so that they may further explore this information and curate their own pathway knowledge. More

generally, we intend on engaging the research community by cultivating partnerships with knowledge database organizations including those that curate information about articles (PubMed), biomolecules (e.g. UniProt, NCBI, ChEBI), pathways and interactions (e.g. Reactome, STRING), model organisms (e.g. Saccharomyces Genome Database (Lang et al., 2018)), and researchers (e.g. ORCID).

To improve the efficiency and utility of Biofactoid, we are developing machine-learning-based NLP technology to support authors in using Biofactoid, as well as to enable the representation of pathways in textual form (Giorgi et al., 2019; Giorgi and Bader, 2020, 2018; Valenzuela-Escárcega et al., 2018). To better accommodate the way individual users prefer to communicate, the system will accept both graphical and textual entry of pathway information as well as automated conversion between these two forms. Assistants will support users to rapidly compose their network, enabling them to add new information from a list of interactions identified in their article by NLP technology (Gyori et al., 2017). Further development of NLP methods for the reliable extraction of pathway information from the publication full-text, combined with development of new tools for curation and quality control, will help realize broad and accurate coverage of pathway knowledge in computable form. We are also improving the search and exploration functions of the system, ensuring Biofactoid information is well connected to other useful knowledge, and that related information is easily accessible starting from a Biofactoid entry. In the future, we envision that information entered by an author in Biofactoid serves as a custom query that can be used to regularly notify the author of new information (e.g. from other publications) related to their interests, such as interacting molecules and phenotypes. This work provides a basis for the development of new technologies to make scientific knowledge more computable and accessible and help researchers identify information within the rapidly growing scientific corpus.

## MATERIALS AND METHODS

### Implementation

Biofactoid is written in JavaScript. The backend server uses a microservice architecture, with Node.js, Express, and RethinkDB. Client-server data synchronization, supporting automatic saving and concurrent editing, uses websockets and a model similar to differential synchronization (Fraser, 2009). The front end uses React and Cytoscape.js (Franz et al., 2016), for network drawing and is optimized for desktop and mobile devices. An administrative dashboard, as well as user curation workflow automation features (e.g. automatic email generation) are integrated into the Biofactoid system to aid system scalability.

### Availability

Biofactoid is available to biomedical researchers for data sharing and exploration on the web at biofactoid.org. To support bioinformaticians and software developers, all user-contributed pathway data is openly accessible in multiple standard formats: JavaScript Object Notation (JSON) for raw data; Systems Biology Graphical Notation Markup Language (SBGN-ML) pathway visualization format using the Process Description notation (Le Novère et al., 2009; van Iersel et al., 2012); and BioPAX (Demir et al., 2010). All code,

documentation and data are open source and freely available through GitHub (github.com/PathwayCommons/factoid); containerized components are freely available on DockerHub (hub.docker.com/r/pathwaycommons/factoid) enabling others to build on and improve the Biofactoid software.

## FUNDING

Biofactoid development was funded by the US National Institutes of Health (NIH) [U41 HG006623, U41 HG003751, R01 HG009979 and P41 GM103504] and the DARPA Big Mechanism and Communicating with Computers programs [ARO W911NF-14-C-0119, W911NF-15-1-0544].

## ACKNOWLEDGEMENTS

We thank Quincey Justman, Miao-Chih Tsai, and Anita DeWaard for feedback on making Biofactoid useful for editors and authors; the Reactome database team for support and feedback on the curation workflow; Alfonso Valencia, and Miguel Vazquez for early support; and the many community members in Toronto, Boston, Portland, and beyond for feedback on the design and concept of Biofactoid.

## COMPETING INTERESTS

None declared.

## REFERENCES

Attwood TK, Agit B, Ellis LBM. 2015. Longevity of Biological Databases. *EMBnet.journal* **21**. doi:10.14806/ej.21.0.803

Bader GD, Cary MP, Sander C. 2006. Pathguide: a pathway resource list. *Nucleic Acids Res* **34**:D504-506. doi:10.1093/nar/gkj126

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The Protein Data Bank. *Nucleic Acids Res* **28**:235–242. doi:10.1093/nar/28.1.235

Bharadwaj A, Singh DP, Ritz A, Tegge AN, Poirel CL, Kraikivski P, Adames N, Luther K, Kale SD, Peccoud J, Tyson JJ, Murali TM. 2017. GraphSpace: stimulating interdisciplinary collaborations in network biology. *Bioinforma Oxf Engl* **33**:3134–3136. doi:10.1093/bioinformatics/btx382

Bornmann L, Mutz R. 2015. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references: Growth Rates of Modern Science: A Bibliometric Analysis Based on the Number of Publications and Cited References. *J Assoc Inf Sci Technol* **66**:2215–2222. doi:10.1002/asi.23329

Brown GR, Hem V, Katz KS, Ovetsky M, Wallin C, Ermolaeva O, Tolstoy I, Tatusova T, Pruitt KD, Maglott DR, Murphy TD. 2015. Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res* **43**:D36-42. doi:10.1093/nar/gku1055

Ceol A, Chatr-Aryamontri A, Licata L, Cesareni G. 2008. Linking entries in protein interaction database to structured text: the FEBS Letters experiment. *FEBS Lett* **582**:1171–1177. doi:10.1016/j.febslet.2008.02.071

Cerami EG, Bader GD, Gross BE, Sander C. 2006. cPath: open source software for collecting, storing, and querying biological pathways. *BMC Bioinformatics* **7**:497. doi:10.1186/1471-2105-7-497

Chinen T, Kannan AK, Levine AG, Fan X, Klein U, Zheng Y, Gasteiger G, Feng Y, Fontenot JD, Rudensky AY. 2016. An essential role for the IL-2 receptor in Treg cell function.

*Nat Immunol* **17**:1322–1333. doi:10.1038/ni.3540

Cordero RJB, de León-Rodriguez CM, Alvarado-Torres JK, Rodriguez AR, Casadevall A. 2016. Life Science's Average Publishable Unit (APU) Has Increased over the Past Two Decades. *PLOS ONE* **11**:e0156983. doi:10.1371/journal.pone.0156983

Demir E, Babur O, Rodchenkov I, Aksoy BA, Fukuda KI, Gross B, Sümer OS, Bader GD, Sander C. 2013. Using biological pathway data with paxtools. *PLoS Comput Biol* **9**:e1003194. doi:10.1371/journal.pcbi.1003194

Demir E, Cary MP, Paley S, Fukuda K, Lemer C, Vastrik I, Wu G, D'Eustachio P, Schaefer C, Luciano J, Schacherer F, Martinez-Flores I, Hu Z, Jimenez-Jacinto V, Joshi-Tope G, Kandasamy K, Lopez-Fuentes AC, Mi H, Pichler E, Rodchenkov I, Splendiani A, Tkachev S, Zucker J, Gopinath G, Rajasimha H, Ramakrishnan R, Shah I, Syed M, Anwar N, Babur O, Blinov M, Brauner E, Corwin D, Donaldson S, Gibbons F, Goldberg R, Hornbeck P, Luna A, Murray-Rust P, Neumann E, Ruebenacker O, Reubenacker O, Samwald M, van Iersel M, Wimalaratne S, Allen K, Braun B, Whirl-Carrillo M, Cheung K-H, Dahlquist K, Finney A, Gillespie M, Glass E, Gong L, Haw R, Honig M, Hubaut O, Kane D, Krupa S, Kutmon M, Leonard J, Marks D, Merberg D, Petri V, Pico A, Ravenscroft D, Ren L, Shah N, Sunshine M, Tang R, Whaley R, Letovksy S, Buetow KH, Rzhetsky A, Schachter V, Sobral BS, Dogrusoz U, McWeeney S, Aladjem M, Birney E, Collado-Vides J, Goto S, Hucka M, Le Novère N, Maltsev N, Pandey A, Thomas P, Wingender E, Karp PD, Sander C, Bader GD. 2010. The BioPAX community standard for pathway data sharing. *Nat Biotechnol* **28**:935–942. doi:10.1038/nbt.1666

Franz M, Lopes CT, Huck G, Dong Y, Sumer O, Bader GD. 2016. Cytoscape.js: a graph theory library for visualisation and analysis. *Bioinforma Oxf Engl* **32**:309–311. doi:10.1093/bioinformatics/btv557

Fraser N. 2009. Differential synchronizationProceedings of the 9th ACM Symposium on Document Engineering - DocEng '09. Presented at the the 9th ACM symposium. Munich, Germany: ACM Press. p. 13. doi:10.1145/1600193.1600198

Gene Ontology Consortium. 2015. Gene Ontology Consortium: going forward. *Nucleic Acids Res* **43**:D1049-1056. doi:10.1093/nar/gku1179

Gerstein M, Seringhaus M, Fields S. 2007. Structured digital abstract makes text mining easy. *Nature* **447**:142–142. doi:10.1038/447142a

Giorgi J, Wang X, Sahar N, Shin WY, Bader GD, Wang B. 2019. End-to-end Named Entity Recognition and Relation Extraction using Pre-trained Language Models. *ArXiv191213415 Cs*.

Giorgi JM, Bader GD. 2020. Towards reliable named entity recognition in the biomedical domain. *Bioinformatics* **36**:280–286. doi:10.1093/bioinformatics/btz504

Giorgi JM, Bader GD. 2018. Transfer learning for biomedical named entity recognition with neural networks. *Bioinforma Oxf Engl* **34**:4087–4094. doi:10.1093/bioinformatics/bty449

Giorgi JM, Nitski O, Bader GD, Wang B. 2020. DeCLUTR: Deep Contrastive Learning for Unsupervised Textual Representations. *ArXiv200603659 Cs*.

Gyori BM, Bachman JA, Subramanian K, Muhlich JL, Galescu L, Sorger PK. 2017. From word models to executable models of signaling networks using automated assembly. *Mol Syst Biol* **13**:954. doi:10.15252/msb.20177651

Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, Turner S, Swainston N, Mendes P, Steinbeck C. 2016. ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res* **44**:D1214-1219. doi:10.1093/nar/gkv1031

Imker HJ. 2018. 25 Years of Molecular Biology Databases: A Study of Proliferation, Impact, and Maintenance. *Front Res Metr Anal* **3**:18. doi:10.3389/frma.2018.00018

Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, Sidiropoulos K, Cook J, Gillespie M, Haw R, Loney F, May B, Milacic M, Rothfels K, Sevilla C, Shamovsky V, Shorser S, Varusai T, Weiser J, Wu G, Stein L, Hermjakob H, D'Eustachio P. 2020. The reactome pathway knowledgebase. *Nucleic Acids Res* **48**:D498–D503. doi:10.1093/nar/gkz1031

Khatri P, Sirota M, Butte AJ. 2012. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput Biol* **8**:e1002375. doi:10.1371/journal.pcbi.1002375

Lang OW, Nash RS, Hellerstedt ST, Engel SR, SGD Project. 2018. An Introduction to the Saccharomyces Genome Database (SGD). *Methods Mol Biol Clifton NJ* **1757**:21–30. doi:10.1007/978-1-4939-7737-6_2

Le Novère N, Hucka M, Mi H, Moodie S, Schreiber F, Sorokin A, Demir E, Wegner K, Aladjem MI, Wimalaratne SM, Bergman FT, Gauges R, Ghazal P, Kawaji H, Li L, Matsuoka Y, Villéger A, Boyd SE, Calzone L, Courtot M, Dogrusoz U, Freeman TC, Funahashi A, Ghosh S, Jouraku A, Kim S, Kolpakov F, Luna A, Sahle S, Schmidt E, Watterson S, Wu G, Goryanin I, Kell DB, Sander C, Sauro H, Snoep JL, Kohn K, Kitano H. 2009. The Systems Biology Graphical Notation. *Nat Biotechnol* **27**:735–741. doi:10.1038/nbt.1558

Liechti R, George N, Götz L, El-Gebali S, Chasapi A, Crespo I, Xenarios I, Lemberger T. 2017. SourceData: a semantic platform for curating and searching figures. *Nat Methods* **14**:1021–1022. doi:10.1038/nmeth.4471

Mack SC, Witt H, Piro RM, Gu L, Zuyderduyn S, Stütz AM, Wang X, Gallo M, Garzia L, Zayne K, Zhang X, Ramaswamy V, Jäger N, Jones DTW, Sill M, Pugh TJ, Ryzhova M, Wani KM, Shih DJH, Head R, Remke M, Bailey SD, Zichner T, Faria CC, Barszczyk M, Stark S, Seker-Cin H, Hutter S, Johann P, Bender S, Hovestadt V, Tzaridis T, Dubuc AM, Northcott PA, Peacock J, Bertrand KC, Agnihotri S, Cavalli FMG, Clarke I, Nethery-Brokx K, Creasy CL, Verma SK, Koster J, Wu X, Yao Y, Milde T, Sin-Chan P, Zuccaro J, Lau L, Pereira S, Castelo-Branco P, Hirst M, Marra MA, Roberts SS, Fults D, Massimi L, Cho YJ, Van Meter T, Grajkowska W, Lach B, Kulozik AE, von Deimling A, Witt O, Scherer SW, Fan X, Muraszko KM, Kool M, Pomeroy SL, Gupta N, Phillips J, Huang A, Tabori U, Hawkins C, Malkin D, Kongkham PN, Weiss WA, Jabado N, Rutka JT, Bouffet E, Korbel JO, Lupien M, Aldape KD, Bader GD, Eils R, Lichter P, Dirks PB, Pfister SM, Korshunov A, Taylor MD. 2014. Epigenomic alterations define lethal CIMP-positive ependymomas of infancy. *Nature* **506**:445–450. doi:10.1038/nature13108

Ostaszewski M, Mazein A, Gillespie ME, Kuperstein I, Niarakis A, Hermjakob H, Pico AR, Willighagen EL, Evelo CT, Hasenauer J, Schreiber F, Dräger A, Demir E, Wolkenhauer O, Furlong LI, Barillot E, Dopazo J, Orta-Resendiz A, Messina F, Valencia A, Funahashi A, Kitano H, Auffray C, Balling R, Schneider R. 2020. COVID-19 Disease Map, building a computational repository of SARS-CoV-2 virus-host interaction mechanisms. *Sci Data* **7**:136. doi:10.1038/s41597-020-0477-8

Pratt D, Chen J, Welker D, Rivas R, Pillich R, Rynkov V, Ono K, Miello C, Hicks L, Szalma S, Stojmirovic A, Dobrin R, Braxenthaler M, Kuentzer J, Demchak B, Ideker T. 2015. NDEx, the Network Data Exchange. *Cell Syst* **1**:302–305. doi:10.1016/j.cels.2015.10.001

Rodchenkov I, Babur O, Luna A, Aksoy BA, Wong JV, Fong D, Franz M, Siper MC, Cheung M, Wrana M, Mistry H, Mosier L, Dlin J, Wen Q, O'Callaghan C, Li W, Elder G, Smith PT, Dallago C, Cerami E, Gross B, Dogrusoz U, Demir E, Bader GD, Sander C. 2020. Pathway Commons 2019 Update: integration, analysis and exploration of pathway data. *Nucleic Acids Res* **48**:D489–D497. doi:10.1093/nar/gkz946

Santos MA, Faryabi RB, Ergen AV, Day AM, Malhowski A, Canela A, Onozawa M, Lee J-E,

Callen E, Gutierrez-Martinez P, Chen H-T, Wong N, Finkel N, Deshpande A, Sharrow S, Rossi DJ, Ito K, Ge K, Aplan PD, Armstrong SA, Nussenzweig A. 2014. DNA-damage-induced differentiation of leukaemic cells as an anti-cancer barrier. *Nature* **514**:107–111. doi:10.1038/nature13483

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**:2498–2504. doi:10.1101/gr.1239303

Slenter DN, Kutmon M, Hanspers K, Riutta A, Windsor J, Nunes N, Mélius J, Cirillo E, Coort SL, Digles D, Ehrhart F, Giesbertz P, Kalafati M, Martens M, Miller R, Nishida K, Rieswijk L, Waagmeester A, Eijssen LMT, Evelo CT, Pico AR, Willighagen EL. 2018. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res* **46**:D661–D667. doi:10.1093/nar/gkx1064

Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, Santos A, Doncheva NT, Roth A, Bork P, Jensen LJ, von Mering C. 2017. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res* **45**:D362–D368. doi:10.1093/nar/gkw937

Todorov PV, Gyori BM, Bachman JA, Sorger PK. 2019. INDRA-IPM: interactive pathway modeling using natural language with automated assembly. *Bioinforma Oxf Engl* **35**:4501–4503. doi:10.1093/bioinformatics/btz289

UniProt Consortium. 2019. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* **47**:D506–D515. doi:10.1093/nar/gky1049

Valenzuela-Escárcega MA, Babur Ö, Hahn-Powell G, Bell D, Hicks T, Noriega-Atala E, Wang X, Surdeanu M, Demir E, Morrison CT. 2018. Large-scale automated machine reading discovers new cancer-driving mechanisms. *Database J Biol Databases Curation* **2018**. doi:10.1093/database/bay098

van Iersel MP, Villéger AC, Czauderna T, Boyd SE, Bergmann FT, Luna A, Demir E, Sorokin A, Dogrusoz U, Matsuoka Y, Funahashi A, Aladjem MI, Mi H, Moodie SL, Kitano H, Le Novère N, Schreiber F. 2012. Software support for SBGN maps: SBGN-ML and LibSBGN. *Bioinforma Oxf Engl* **28**:2016–2021. doi:10.1093/bioinformatics/bts270

Waagmeester A, Stupp G, Burgstaller-Muehlbacher S, Good BM, Griffith M, Griffith OL, Hanspers K, Hermjakob H, Hudson TS, Hybiske K, Keating SM, Manske M, Mayers M, Mietchen D, Mitraka E, Pico AR, Putman T, Riutta A, Queralt-Rosinach N, Schriml LM, Shafee T, Slenter D, Stephan R, Thornton K, Tsueng G, Tu R, Ul-Hasan S, Willighagen E, Wu C, Su AI. 2020. Wikidata as a knowledge graph for the life sciences. *eLife* **9**. doi:10.7554/eLife.52614

Wang G, Meyer JG, Cai W, Softic S, Li ME, Verdin E, Newgard C, Schilling B, Kahn CR. 2019. Regulation of UCP1 and Mitochondrial Metabolism in Brown Adipose Tissue by Reversible Succinylation. *Mol Cell* **74**:844-857.e7. doi:10.1016/j.molcel.2019.03.021

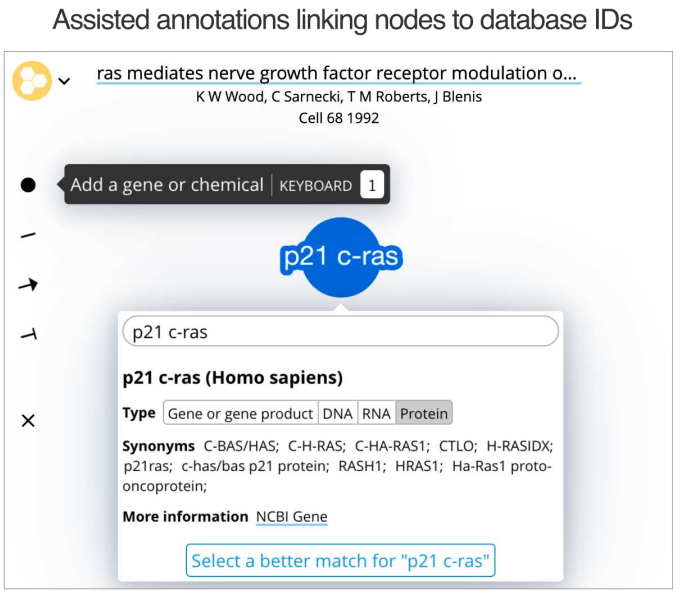
Wang T, Cao Y, Zheng Q, Tu J, Zhou W, He J, Zhong J, Chen Y, Wang J, Cai R, Zuo Y, Wei B, Fan Q, Yang J, Wu Y, Yi J, Li D, Liu M, Wang C, Zhou A, Li Y, Wu X, Yang W, Chin YE, Chen G, Cheng J. 2019. SENP1-Sirt3 Signaling Controls Mitochondrial Protein Acetylation and Metabolism. *Mol Cell* **75**:823-834.e5. doi:10.1016/j.molcel.2019.06.008

Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Geer LY, Helmberg W, Kapustin Y, Kenton DL, Khovayko O, Lipman DJ, Madden TL, Maglott DR, Ostell J, Pruitt KD, Schuler GD, Schriml LM, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Suzek TO, Tatusov R, Tatusova TA, Wagner L, Yaschenko E. 2006. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **34**:D173-180. doi:10.1093/nar/gkj158

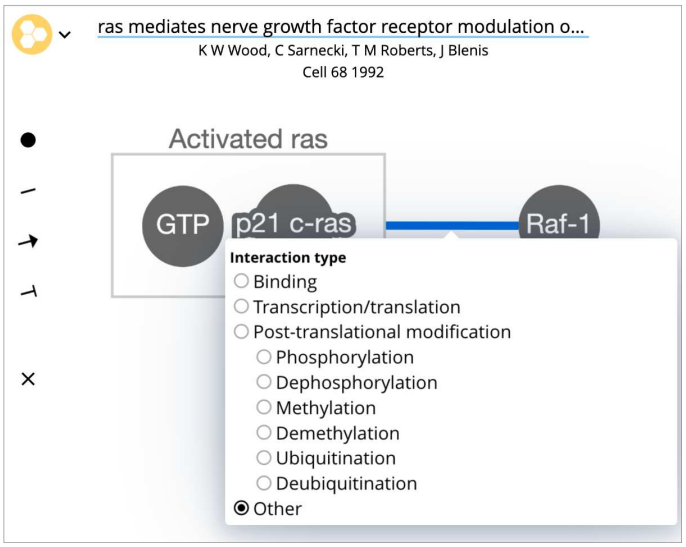
# FIGURES

## Figure 1

### A



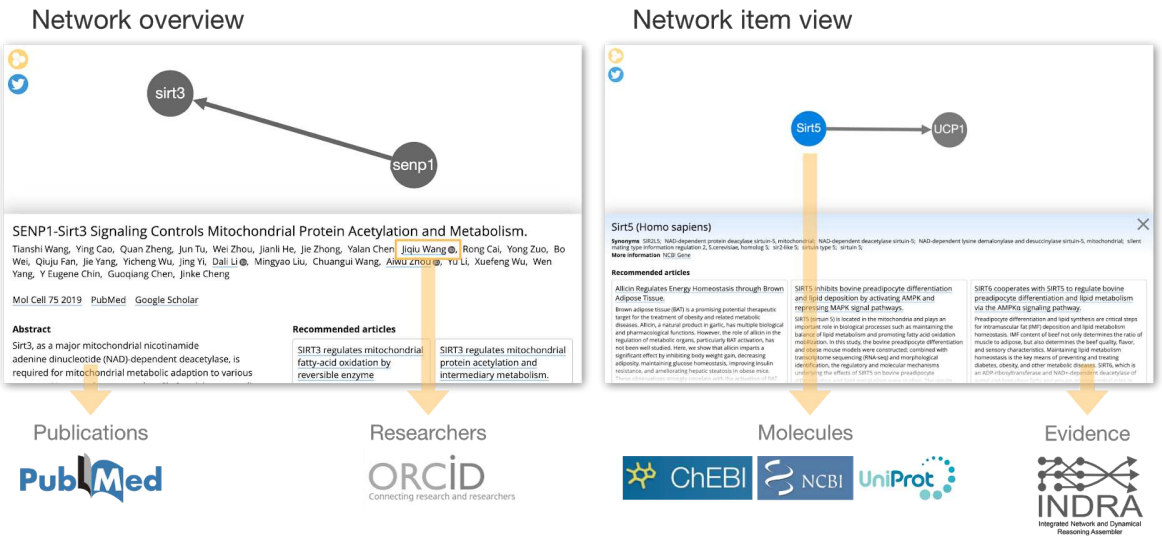

### B



# Figure 2

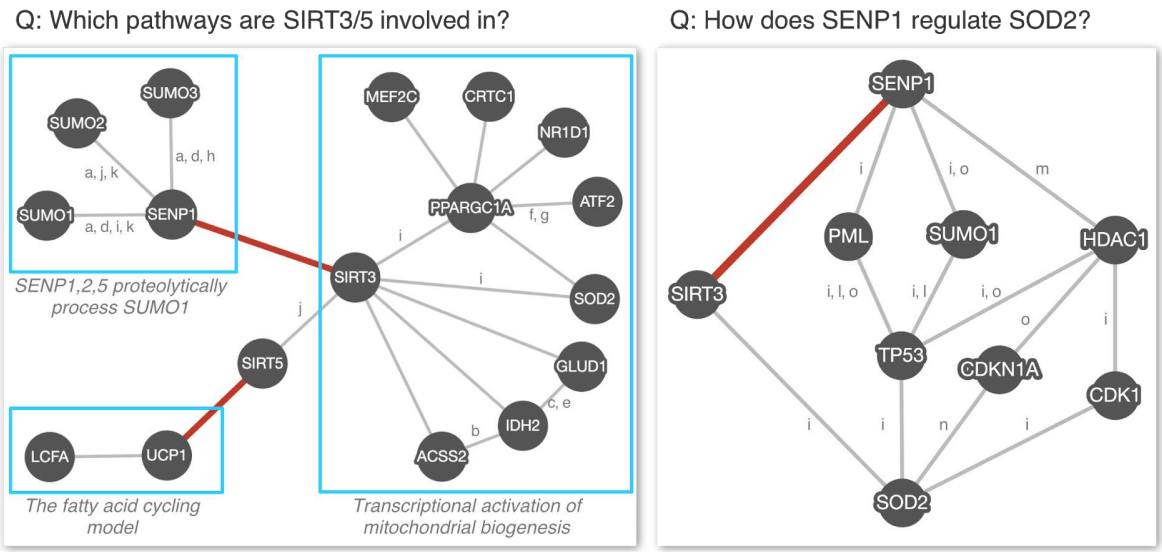
## A

### Research connected to information sources




## B

### Powerful searches across data sources




Reactome Pathways
Biofactoid
Pathway Commons

a. IntAct, b. Recon X, c. HumanCyc d. BioGRID, e. Pathbank, f. Molecular Signatures Database, g. NCI Pathway Interaction Database: Pathway, h. Biomolecular Interaction Network Database, i. Comparative Toxicogenomics Database, j. BioGRID, k. Database of Interacting Proteins, l. PANTHER Pathway, m. NetPath, n. PhosphoSitePlus, o. Reactome

**Figure 3**

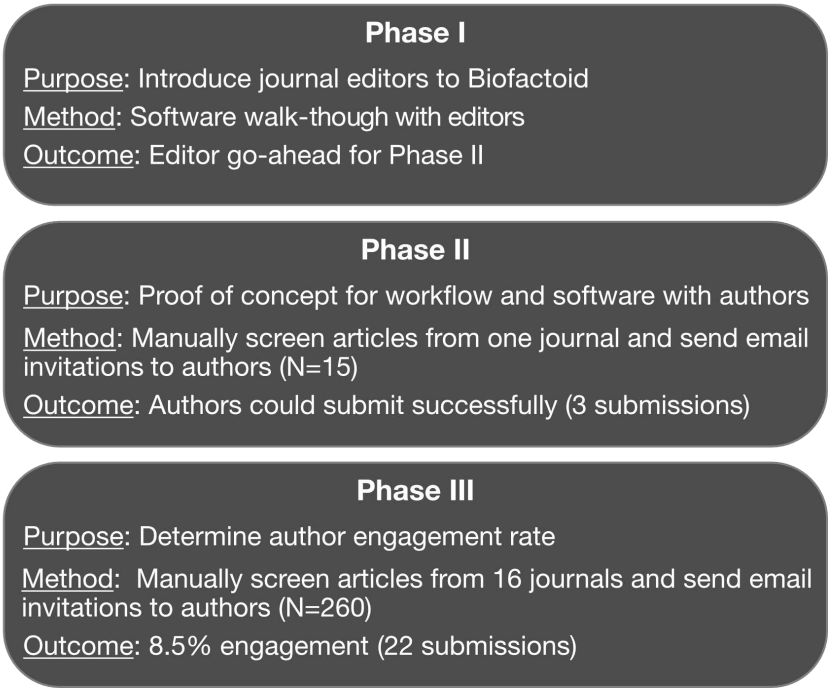


**Figure 1. The Biofactoid curation tool.** Curation in Biofactoid involves drawing a network of relationships between genes or molecules. **(A)** Genes and chemicals are represented by circles (nodes) where users provide a label, the type of gene product, and the organism. A custom search engine matches the label to a corresponding record from a database of genes or chemicals. **(B)** Relationships are represented by connecting participants with lines, arrows (to indicate activation), or 'T-bars' (to represent repression). Users select the mechanism that best describes the interaction. Complexes are represented as participants enclosed by a box.

**Figure 2. Biofactoid data is connected to existing knowledge and enables more powerful search. (A)** The Biofactoid Explorer is an interactive web app that publicly presents each author-curated entry alongside their article. Arrows indicate how curated information is connected to outside knowledge bases. A "Network overview" (left) displays information about the article and pathway as a whole; a "Network item view" (right) displays information for a selected item (e.g. interaction, protein). **(B)** Biofactoid data integrated with existing structured biological knowledge enables powerful searches across data sources. Two author-curated interactions submitted to Biofactoid (red edges) bridge previously distinct pathways from the Reactome Pathway Database involved in mitochondrial biogenesis (left) and provide a new, more direct regulatory route between two mitochondrial genes (right). Pathway and interaction information was provided by Pathway Commons (pathwaycommons.org), a web resource that provides a single point of access for multiple public interaction and pathway databases.

**Figure 3**. **Biofactoid pilot study.** A three-phase pilot tested the feasibility of Biofactoid and involved journal editors and authors of research articles. Phase I and II involved editors and authors whose articles were recently published. In Phase III, 2065 published articles were screened and authors of suitable articles were invited to Biofactoid. The articles screened were from 16 journals (Table S1).

## SUPPLEMENTARY FIGURES AND TABLES

**Table S1. Prevalence of articles with pathway knowledge suitable for Biofactoid**

| ISSN | [1]Journal | [2]Coverage [Vol(Iss)] | Articles screened | Hits | % Hits |
|---|---|---|---|---|---|
| 2211-1247 | Cell Reports | 30(1) - 32(11) | 953 | 109 | 10.3 |
| 1097-4164 | Molecular Cell | 73(1) - 79(6) | 725 | 85 | 10.5 |
| 1549-5477 | Genes & Development | 34(1-2) - 34(17-18) | 93 | 15 | 13.9 |
| 1476-4679 | Nature Cell Biology | 22(4) - 22(9) | 84 | 10 | 10.6 |
| 1083-351X | Journal of Biological Chemistry | 295(31) - 295(37) | 210 | 21 | 9.1 |
| **Total** | - | - | 2065 | 240 | - |
| **Weighted Average** | - | - | - | - | 10.4 |

[1]Only journals in which at least 80 'Hits' were identified were included. A 'Hit' is an article that provides direct evidence for a molecular interaction that can be captured by Biofactoid. The ability of Biofactoid to capture an interaction depends upon the type of bioentities, the relationship types and organisms described in the article. In total, articles from 16 journals were screened: EMBO; Molecular and Cellular Biology; Cell; Cancer Cell; iScience; Cell Metabolism; Science; Nature Genetics; Science Signaling; Science Advances; Immunity; Cell Reports; Molecular Cell; Genes & Development; Nature Cell Biology. [2]Coverage indicates the span of journal issues that were included. Only primary research articles from each issue were screened.