Information Research - Vol. 30 No. iConf (2025)

The role of ontologies in machine learning: a case study of gene ontology

Qiaoyi Liu and Jian Qin

DOI: https://doi.org/10.47989/ir30iConf47575

Abstract

Introduction. Ontologies as knowledgebases have been heavily applied in computational biological studies by implementing into ML models for purposes such as disease-gene associations identification.

Method. We conduct a case study using gene ontology (GO) annotation data and three ML models to replicate the prediction of autism spectrum disorder (ASD)-causing genes.

Analysis. Data were collected from GO and Simmons Foundation Autism Research Initiative (SFARI). The semantic similarities between GO annotation terms on gene products were calculated.

Results. The best-performing model can reach an AUC of .85, which means using GO annotation data for ASD disease-gene prediction can receive a significantly accurate result. However, we stress the importance of constructing knowledgebases in adapting to LLMs and the role of LIS professionals in curating community knowledge for interoperability and reuse.

Conclusion. Biomedical ontologies play a crucial role in the discovery of biomedical knowledge. Knowledge organization and computer science domains require more communication and synchronization in the face of emerging AI and ML technologies.

Introduction

Ontologies as knowledge representation and organization systems have had a prolific increase in the last thirty years. This is particularly the case in the biomedical field. As of this writing, BioPortal, the world's most comprehensive repository of biomedical ontologies, have registered 1,145 ontologies that contain a total of 15.6 million classes, 36,286 properties, and 99.5 million mappings (National Center for Biomedical Ontology, 2024). The vast amount of structured data resulted from these ontologies have been utilized by researchers to explore disease-gene relations and predict genes that have high risk to specific diseases such as Alzheimer's disease (Asif et al., 2018) (Yang et al., 2021) (Binder et al., 2022). The fast growth of biomedical data with parallel advancement of machine learning (ML) algorithms and computational tools created the necessary condition for scientists to respond to urging problems such as identification of genomics for disease research (Piñero et al., 2020) (Krishnan et al., 2016), protein structure analysis (Radivojac et al., 2008), phylogenetic inference (Ata et al., 2021).

The well-structured and actively curated data in knowledge organization systems provide the quality and structures desired by ML model developers, especially in the shift from model-centric artificial intelligence (AI) to data-centric AI (Ng, 2024). However, there has been a lack of communication and understanding between different communities of information science, computer science, related scientific disciplines on the role of knowledge organization systems (KOS), the work involved in building them, and how KOS may be transformed into knowledgebases for AI modelling and applications. The work involved in building KOS includes not only defining subject terms and mapping controlled terms to free-text keywords, but more importantly, defining semantic relations between entities/concepts, bridging raw and machine-readable data, and curating and organizing structured and interoperable data through human intervention and/or automatic methods. This type of knowledge work is essential in bringing scattered, natural language information into systematic representations of knowledge, i.e. data with interpretations (Haendel et al., 2018) for machine to process and even conduct automated reasoning. Despite ontologies and knowledgebases are being constructed across disciplinary fields, the theoretical and methodological aspects of these knowledge organization and representation systems have been largely confined in separate communities, which hinder the communication and sharing of research across communities who study and build KOS.

The purpose of this paper is to elucidate the role of KOS as a trustable data source for AI/ML modelling and applications. We will start with reviewing the current developments of ontologies and knowledgebases, highlighting how knowledgebase structure and design can improve its functions in a ML research workflow. Using a case study of gene ontology (GO), we will illustrate the role that gene ontology played in building ML prediction models. In this case study the GO annotation data and supervised ML models are used to calculate the functional similarities of autism syndrome disease (ASD) genes. The detail of methods and results of the case study is presented in Section 3, followed by Sections 4 and 5, where we speculate on the practices and factors contributing to the quality of GO annotation data. Future KOS research direction is discussed on processing multiple data formats, e.g., images and data sources, disease ontologies, and clinical database, to accommodate AI/ML models.

Evolving knowledge organization systems

There are four main types of KOS. Based on the level of sophistication, the simplest type is term lists, such as glossaries and dictionaries. The second type is metadata-like models, including gazetteers, directories, and authority files. Classification and categorization go to the next level of sophistication because of the embedded relations between concepts or classes in classification/categorization schemes, taxonomies, and subject headings. The relationship models are the most sophisticated among the four KOS types. Ontologies and semantic networks are the two members of the relationship model, which possess all traits of a KOS that one can hope for: explicitly representing concepts/entities with unambiguous terms while embedding relations between concepts/entities (Zeng, 2008). In the sense of representing a conceptual system via a logical theory, an ontology consists of an annotated and indexed set of formal propositions or assertions about things, a collection of assertions that are called a theory in logic (Guarino & Giaretta, 1995). This special property of ontologies is as closely as it can be in fitting the knowledge representation ideology in AI, which emphasizes sufficiently precise notation, adequacy and expressiveness of representation schemes (Bench-Capon, 1990).

Ontologies sometimes are also called knowledgebases because they are essentially a collection of symbolic structures representing the world based on our cognition (Levesque & Lakemeyer, 2022). Such representation is semantically rich with not only unambiguous vocabularies for entities and individuals in the entities, but also explicit relations between the entities that go beyond hierarchical and associative relations that are the only available relation types in many traditional KOS. Not all ontologies, however, can be called knowledgebases. For example, Schema.org is an ontology developed in collaboration among Google, Microsoft, Yahoo, and Yandex to be used for representing web content as structured data (Google et al., 2024). The purpose of Schema.org is to provide a standard representation scheme for classes of entities and properties these entities possess. Since it is just a representation scheme, it does not contain data (i.e., individual members of classes). Therefore, it is not a knowledgebase and cannot be used for modelling or reasoning purposes.

While many traditional KOS are not designed for problem-solving nor as a data source for ML modelling purposes, many ontologies in the biomedical domain have disrupted the tradition and evolved into knowledgebases. For instance, gene ontology (GO) (Ashburner et al., 2000) and disease ontology (DO) (Bello et al., 2018) use axioms to model knowledge in order to define the semantic interpretation of the presented entities, rules, and class constraints and present multiple relations between concepts. Knowledge organization and AI communities, two research fields that were once separated and did not have much communication before, are now drawing closer and converging through advances in semantic web technologies and data science (Qin, 2020). This trend symbolizes a change to knowledge organization practices. More importantly, it is a signal to knowledge organization moving towards a more interoperable, practical, and heterogeneous identity that exceeds the simple purpose of storing knowledge.

One application of biomedical ontologies is in disease-gene association discovery and identification. Advanced genome sequencing technologies accelerated the process of exploring genomic variations (Piñero et al., 2020) and genetic markers' detection (Chang et al., 2024), generating vast amount of data. This data is preserved and organized into KOS by biocurators who apply knowledge organization theories and practices, which creates semantically rich data sources for applying ML algorithms to analyse larger and complex data sets (Libbrecht & Noble, 2015). Complex diseases with a strong genetic influence often have multiple aetiologies with the involvement of possibly hundreds of different genes. Supervised ML methods can trace hidden relationships among disease-causing genes in existing datasets to discriminate unknown disease genes from non-disease genes (Asif et al., 2018). This advancement plays a crucial role in disease diagnosis and form the basis for clinical decision-making (Chang et al., 2024). This approach was once difficult to proceed due to the lack of: (a) structured KOS with semantically rich relationships between properties; and (b) a powerful computational hardware with feasible models to analyse large heterogeneous data sets. Problem (a) was greatly improved through making data FAIR (findable, accessible, interoperable, reusable) (Wilkinson et al., 2016) and encoding languages (e.g. XML and JSON) for representing machine-readable knowledge and reflecting reality. Problem (b) now is being addressed by numerous supervised and unsupervised ML models that are developed and applied to assist in areas such as genetics and molecular science.

Case study: using GO annotation and ML models to identify autism spectrum disorder (ASD)

Case selection

Gene ontology (GO)

The fast-developing nucleotide sequencing techniques and gene expression analysis has urged the biological community to establish a knowledge resource for this massive data. Unlike other STEM research fields, biological knowledge can be less explained by mathematical equations but more through natural language (Pesquita et al., 2009). Gene ontology (GO, http://geneontology.org) constructed by the gene ontology consortium, is crowned as the GOld mine. It provides "a comprehensive, structured, computer-accessible representation of gene function for genes from any cellular organism or virus." Until 2023, GO contains 43,303 biological terms as annotations to gene products, linked together by 88,099 relationships (The Gene Consortium, 2023). The GO knowledgebase represents a standardized controlled vocabulary which defines various components of molecular biology shared amongst life forms (Yousef et al., 2021). It has become a critical component of life science research, supporting analysis of large-scale genomics data analysis and biological systems (Duck et al., 2016) and broadly used in research, clinical diagnosis, and industry. Over the years many ML models and equations were developed specifically for processing and using GO data. Considering its significance to biological research and established computational techniques, we select GO annotation data as our primary data source to conduct case study.

GO construction and GO annotations

Bio-entities described in GO can be considered as knowledge unites, or 'things' such as gene products. The entities in GO are structured as a directed acyclic graph (DAG) in which GO terms/annotations are represented as nodes and relationships between terms are represented as edges that follow certain directions and never form a closed loop (Asif et al., 2018). Each ontology term (called 'class' in the field of ontology) represents a functional characteristic that can be attributed to a gene product (The Gene Consortium, 2023). Terms representing the 'things' are organized into three categories: molecular function (MF), biological process (BP), and cellular component (CC). Each term is described by five required elements: a unique ID, term name, aspect (which category it belongs to), definition, and relationships to other terms. GO links the terms by using a set of triple statements, most commonly '*is_a*' or '*part_of*', which stands for class-subclass relationship and part-whole relationship, respectively (see table 1). A GO annotation is an association between a specific gene product and a GO term and should be interpreted as a statement that the gene product possesses the functional characteristics represented by the GO term. Each GO annotation covers only one characteristic of the gene product. Therefore, a gene product can have multiple GO annotations. GO annotations are continually added to the knowledgebases from 173,000 scientific papers. All annotations are supported by an evidence code which describes the type of evidence and a reference that lists a persistent identifier for tracing the source. For quality control, they are regularly reviewed, edited, or removed by biocurators or the GO user community (The Gene Consortium, 2023).

Relation	Description	Example
is_a	The basic structure of GO. If we say A *is_a* B, we mean that entity A is a subtype of entity B.	Mitotic cell cycle *is_a* cell cycle, or lyase activity *is_a* catalytic activity.
part_of	The *part of* relation is used to represent part-whole relation. A *part of* relation would only be added between A and B if B is **necessarily** *part of* A: wherever B exists, it is as *part of* A, and the presence of the B implies the presence of A.	If a gene product X is annotated as located in the inner mitochondrial membrane and the ontology records a *part of* relation between inner mitochondrial membrane and mitochondrion, we can safely conclude that X is located in a mitochondrion.
has_part	The logical complement to the *part of* relation is *has part*, which represents a part-whole relationship from the perspective of the parent. As with *part of*, the GO relation *has part* is only used in cases where A always has B as a part, i.e., where A necessarily *has part* B. If A exists, B will always exist; however, if B exists, we cannot say for certain that A exists. i.e., all A have part B; some B part of A.	A receptor tyrosine kinase activity *has part* ATP hydrolysis activity. However, it would not then be correct to group all annotations to kinase activity under ATPase activity.
regulates	A relation that describes case in which one process directly affects the manifestation of another process or quality, i.e., the former *regulates* the latter. The target of the regulation may be another process, for e.g., regulation of a pathway or an enzymatic reaction, or it may be a quality, such as cell size or pH. Analogously to *part of*, this relation is used specifically to mean necessarily *regulates*: if both A and B are present, B always *regulates* A, but A may not always be regulated by B., i.e., all B *regulate* A; some A are *regulated by* B.	If gene product X is annotated as involved in a process that *regulates* glycolysis, it would not be correct to conclude that X participates in glycolysis.

Table 1. Main term relations used in GO

Identify disease-gene associations in ASD

Disease-gene association prediction has been in the spotlight of bioinformatics research for over a decade. Scholars and doctors are eager to identify gene mutations, which are the primary causes of genetic diseases. Many large-scale genetic studies provided candidate genes that may cause diseases. However, traditional disease-gene association costs heavy human labor and are highly difficult to process due to genetic heterogeneity. The disease-causing genes are identified by statistical geneticists, where linkage analysis and association studies were conducted on candidate genes based on their likelihood of being involved in a specific disease i.e. gene prioritization algorithms (Radivojac et al., 2008). With more heterogeneous data, it is no longer possible to conduct this type of analysis solely by human labor.

Nowadays, computational approaches can speed up this process and are capable of handling more complex diseases, such as autism spectrum disorder (ASD). ASD is a type of neurodevelopmental syndrome that affects about one in 100 people worldwide (Geschwind & State, 2015). Strong evidence indicates the causes include both genetic and environmental factors (Kim & Leventhal, 2015). Diseases like ASD tend to present a highly heterogeneous genotype, which makes it difficult for biological marker identification. Although ML methods can be used to identify these markers, their performance highly depends on the size and quality of available data (Asif et al., 2018). Therefore, using GO annotation data can provide consistent and clean data for ML models to learn from, thus improving their general performance. Our purpose is not developing a new model or equation, but to replicate this process and discuss the data quality and ontology construction from the LIS scope outside of these two communities. We hope to highlight the significance of structured knowledgebases for computational biological research and applicability in AI LLMs.

Functional similarity vs. semantic similarity

One major issue in identifying disease-gene association is how to define an unknown gene can cause similar outcomes as a disease-associated gene already known. Biologists found that functional similar genes tend to contribute to similar phenotypes. For instance, etiologically relevant genes disrupted by genetic variants in ASD patients tend to aggregate in specific biological processes (Voineagu & Eapen, 2013). This means disease-causing genes and disease-candidate genes may belong to the same tree path in GO DAG structure (Asif et al., 2018). Gene products that share highly overlapping GO terms may have higher functional similarities. If a gene product is on the same tree branch with gene products that are associated with disease-related GO annotation function terms, it indicates this gene may have a higher probability to be a disease-causing gene. Thus, the comparison of gene products' function similarities is transformed into the comparison of semantic similarities of GO terms. Adopting ontology annotation data can provide a means to compare entities on aspects that otherwise not be comparable.

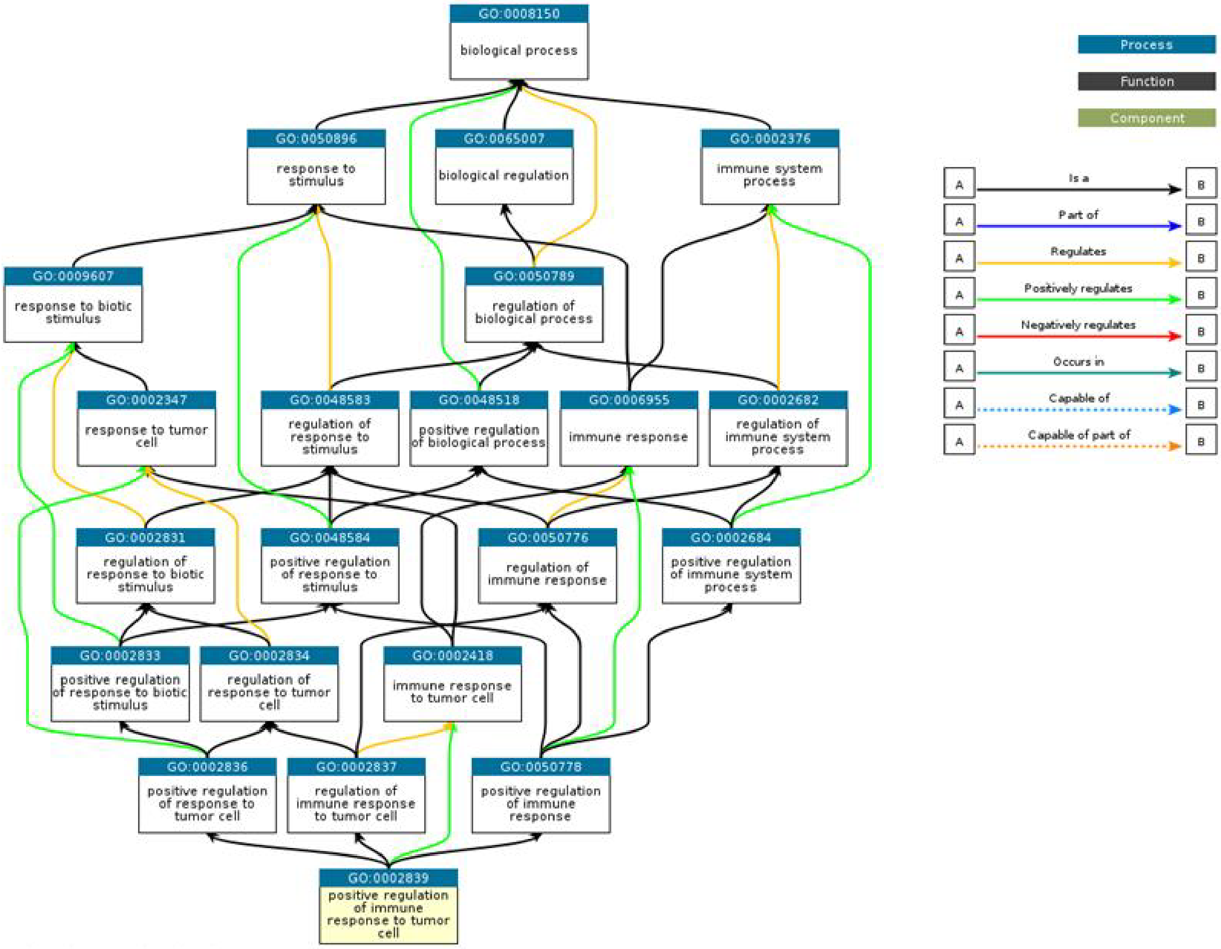
Typically there are two ways to compare terms in graph-structured ontologies such as GO: edge-based or node-based (Pesquita et al., 2009). Edge-based approaches are based on counting the number of edges in the graph path between two terms (Rada et al., 1989). This can be problematic for biological data because the approach is based on two conditions: (i) nodes and edges in the biological ontology are uniformly distributed, and (ii) edges at the same level in the ontology correspond to the same semantic distance between terms. However, biological knowledge can rarely meet these two conditions where terms at the same tree structure level share the same scale and weight.

Node-based approaches are more commonly accepted in biological domain. There have been several statistical measurements and equations developed. Resnik proposed the Information Content (IC) to quantify the informativeness of a concept c as negative the log likelihood (Resnik, 1999):

$$-\log p(c) \tag{1}$$

However, a drawback of Resnik's method is that it ignores the information contained in the structure of the ontology by only concentrating on the information content of a term derived from the corpus statistics. For biological ontologies, the specificity of a GO term is usually determined by its location in the GO graph. A GO term's semantics (biological meanings) are inherited from all its ancestor terms (Wang et al., 2007). For instance, shown in figure 1, GO:0002839 positive regulation of immune response to tumour cell is a child term of GO:0002418 immune response to tumour cell, the latter is a child term of GO:0002347 response to tumour cell and GO:0006955 immune response, both of which are the child terms of GO:0008150 biological processes. Because GO term is the aggregation of all its parent terms, GO:0002839 positive regulation of immune response to tumour cell should possess the characteristics of both GO:0002418 response to tumour cell and GO:0006955 immune response. Not to mention biomedical ontologies usually have various edge length i.e. edges at the same level convey different semantic distances, various depth i.e. terms at the same level have different level of details, and various node density i.e. some areas of the ontology have a greater density of terms than others (Pesquita et al., 2009).

Wang proposed a metrics that is specifically for encoding biological terms' semantics by aggregating the semantic contributions of all its ancestor terms including itself in the GO graph. Wang believes Resnik's IC approach focus is more applicable for knowledge in natural language such as bird and crane, forest and graveyard but is not the ideal option for biological knowledge in ontologies (Wang et al., 2007). IC neglects the logic that if two GO terms share the same parent are near the root of the ontology i.e., terms that are more general, they should have larger semantic difference than two terms having the same parent and being far away from the root of the ontology because the latter are more specific terms. GO is constructed in such way that if the child GO term describes the gene product, then all its parent terms must also apply to that gene product (Wang et al., 2007). DO contains logical definitions (axioms) to describe relevant disease drivers, constructed with specific relational ontology (RO) terms, to create a restriction between a DO term and another open biological and biomedical ontology (OBO) Foundry ontology term. Thus, DO has the ability to 'infer' the child terms originated from a parent term based on a known parent-child relationship (Qin & Liu, 2024). Therefore, it is reasonable to aggregate the biological meanings of all its ancestor terms when determining its semantic similarity with another GO term.


GO:0008150
biological process
GO:0050896
response to stimulus
GO:0065007
biological regulation
GO:0002376
immune system process
GO:0009607
response to biotic stimulus
GO:0050789
regulation of biological process
GO:0002347
response to tumor cell
GO:0048583
regulation of response to stimulus
GO:0048518
positive regulation of biological process
GO:0006955
immune response
GO:0002682
regulation of immune system process
GO:0002831
regulation of response to biotic stimulus
GO:0048584
positive regulation of response to stimulus
GO:0050776
regulation of immune response
GO:0002684
positive regulation of immune system process
GO:0002833
positive regulation of response to biotic stimulus
GO:0002834
regulation of response to tumor cell
GO:0002418
immune response to tumor cell
GO:0002836
positive regulation of response to tumor cell
GO:0002837
regulation of immune response to tumor cell
GO:0050778
positive regulation of immune response
GO:0002839
positive regulation of immune response to tumor cell
Process
Function
Component
A
Is a
B
A
Part of
B
A
Regulates
B
A
Positively regulates
B
A
Negatively regulates
B
A
Occurs in
B
A
Capable of
B
A
Capable of part of
B

QuickGO - https://www.ebi.ac.uk/QuickGO

Figure 1. A graphic view of GO:0002939 positive regulation of immune response to tumour cell and all its ancestor terms in GO

Next, the issue is transforming semantic similarity of GO terms into the functional similarity of gene products. Since one gene product may have more than one GO annotation, we must consider the contributions from the semantically similar terms that annotate the genes separately (Wang et al., 2007). Wang first defines the maximum semantic similarity between one GO term and a set of GO = $\{go_1, go_2, \dots, go_k\}$as:

$$Sim(go, GO) = \max_{1\le i\le k}(S_{GO}(go, go_i)) \tag{1}$$

Then, given two gene products G1 and G2 annotated by sets of GO terms: GO1 = $\{go_{11}, go_{12}, \dots, go_{1m}\}$ and GO2 = $\{go_{21}, go_{22}, \dots, go_{2n}\}$ respectively, the functional similarity of G1 and G2 can be defined as:

$$Sim(G_1, G_2) = \frac{\sum_{1 \le j \le m} Sim(go_{1j}, GO_2) + \sum_{1 \le j \le n} Sim(go_{2j}, GO_1)}{m+n} \quad (2)$$

In this study, we applied Wang's approach to calculate the functional similarity of gene products for predicting candidate genes that lead to ASD disease. GO annotation data were collected using the *org.Hs.eg.db* R package (Carlson et al., 2019). Only terms for biological processes are selected (N = 28,140). Next, we mapped all the GO terms and their associated gene products (N = 157,247). Since our aim is investigating the construction and quality of GO as a knowledge ontology in biomedical research, we only try to replicate part of the workflow by Asif et al., using three of the ML models – support-vector machines (SVM), random forest (RF), and gradient boosting (GB). We randomly selected 1,000 gene products from the mapping as our train set for the models. For our test set, we also randomly selected 20 candidate gene products obtained from the Simons foundation autism research initiative (SFARI, https://gene.sfari.org/) gene database (N = 1,176), in which 15 are categorized by SFARI as high confidence disease genes (HD) and 5 are categorized as low confidence (LD) disease genes. This categorization is used to compare with results from the ML classification models. Table 2 shows the 20 test gene products and their categories. Before we allow the model to conduct semantics similarity calculation, we removed SFARI's categorization for the models to predict and then verify their overall performance by comparing its results with SFARI's classification. The semantic similarity measures between GO terms were implemented using the *GOSemSim* R package (Yu et al., 2010).

Results

As is shown in Fig. 2, we used the train set gene product and their GO annotation data to run the 1,000×1000 similarity matrix. Next, we conducted the 1000×20 similarity matrix between train set and test set (see figure 3). Based on their functional similarity, we identify which of the 20 test set gene products may be a disease-associated gene. Out of the three models we used, SVC-based classifier trained and test on Wang's semantic similarity matrix outperformed the other classifiers, with AUC value equals 0.85 (see table 3). The difference between RF and GB AUC values were minor, indicating the independence of the methodology to the semantic measure. A recall of 0.25 is considerably low. However, given the highly imbalanced dataset (most genes in train set are non-ASD-associated with only a very small number of genes are ASD-associated) and small number of train dataset used, we believe this is an expected and sufficient result to prove that GO annotation data could be used in gene-disease association identification ML models. Furthermore, the overall performance of MLs is crucially dependent on the quality of GO annotation data and structural relationship between GO terms. Here we only demonstrated one way to calculate the semantic similarity between GO terms. The results can vary depending on which type of approach one applies.

Gene product	Candidate gene category by SFARI (ASD causing is 1, non-ASD causing is 0)
SNTG2	1
BIRC6	1
CNTN4	1
ADORA2A	1
LZTR1	1
WWOX	1
HYDIN	1
RBFOX1	1
TRPM1	1
FAN1	1
CMPK1	1
STIL	1
TAL1	1
BCL9	1
OR4A47	1
CARD16	0
CASP1	0
PCDH17	0
P2RX6	0
AHR	0

Table 2. Selected test set gene products and ASD categories (N=20)

Classifiers	AUC	Recall	F1 Score
SVC	0.85	0.25	0.10
RF	0.43	0.25	0.10
GB	0.50	0.25	0.10

Table 3. The performance of classifiers trained and tested over Wang semantic similarities matric. The Area Under the Curve (AUC) evaluation metric was used to estimate and compare the performance of the classifiers

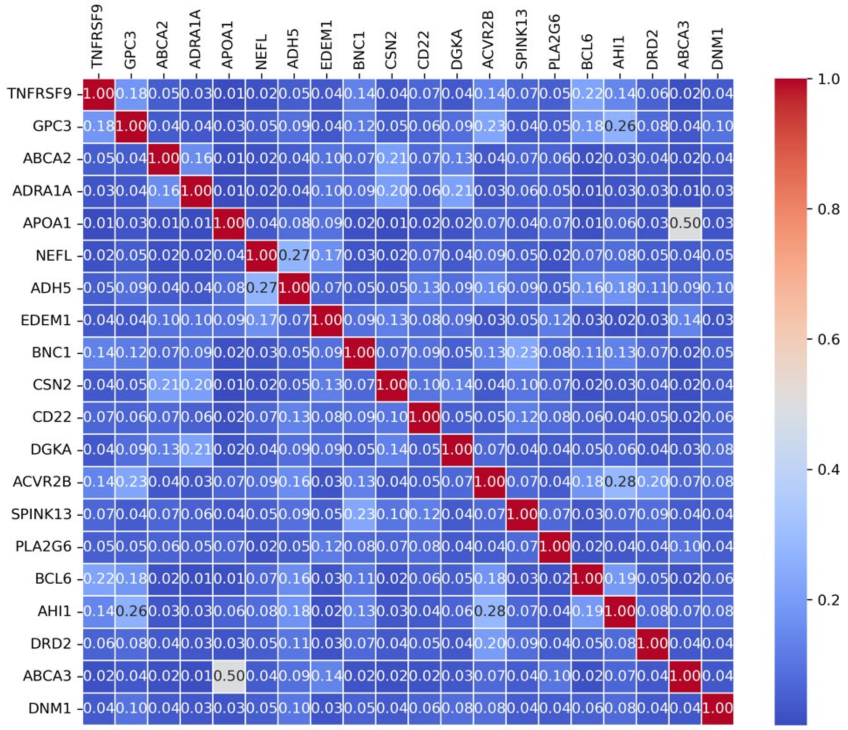


Figure 2. Similarity matrix of the first 20 gene products from train set

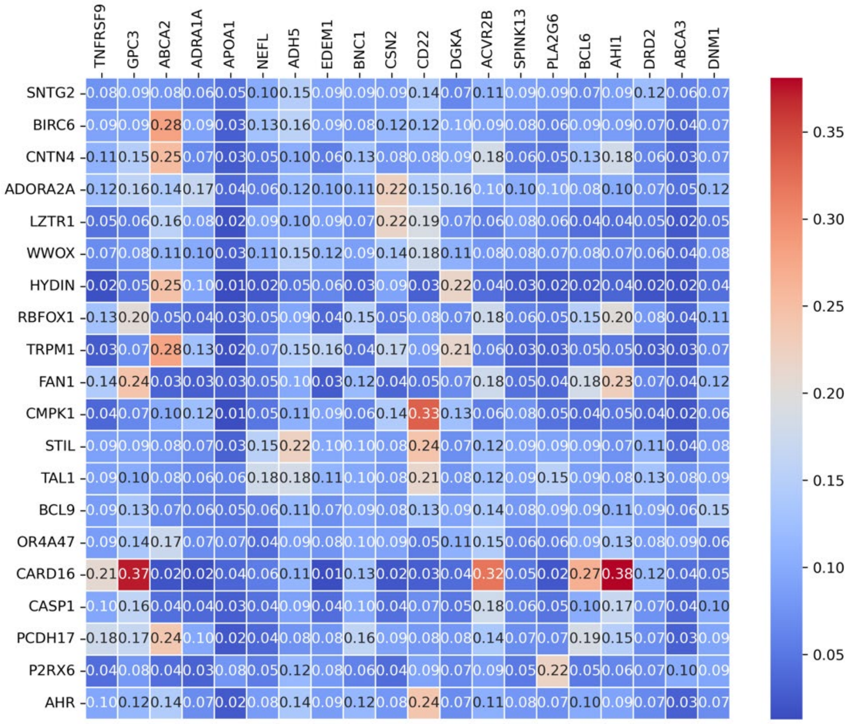


Figure 3. Similarity matrix of the first 20 gene products from train set and all 20 gene products from test set

One major issue which reduces its performance is the imbalance data between one class from other classes. For example, in clinical or disease-related cases, there is inevitably less data from treatment groups than from the normal (control) group (Min et al., 2016). One of the common solutions to this problem is data pre-processing. This is largely conducted by data curators guided by schemas and controlled vocabularies. However, the high proportion of duplicate or near-duplicate samples in biological sequencing data is a serious problem during ML model training which tends to be overlooked (Auslander et al., 2021). Careful data processing is needed to ensure the independence of data between train and test set. When using GO annotation data to quantify the similarity between biological terms, an important issue is that some annotation data in GO is referred from other sources based on similarity. Using these annotations for semantic similarity calculation is in fact, data circularity (Pesquita et al., 2009).

Discussion

Conventional LIS approaches to knowledge representation such as hierarchical and faceted classification integrated human knowledge into a systematic arrangement in which concepts and structures tend to be abstract and have an epistemology orientation. Ontologies, on the contrary, disintegrate parts of knowledge into a problem-solving focused structure and are more pragmatic and application oriented. The disintegrative approach places the entity itself, in our case it is the gene products, in a less important position, but rather, focuses on all aspects related to it i.e., GO terms and annotations (Qin, 2002). As a result, ontologies are more suitable as conceptual frameworks to specific problems. In this case study scenario, GO annotation can be applied to various statistical measurements of semantic similarity, which represents the entity i.e., gene product it describes. The application of GO knowledgebase has transcended beyond its initial purposes which is representing and organizing a phenomenon of a knowledge domain.

As Deep Learning and LLM models are more frequently applied to bioinformatics and biomedical research, we must reconsider schema and frameworks for building knowledge organization systems in order for them to apply to more complex computational approach. Already there have been ways to use LLMs in knowledge graph engineering and ontology construction in replacement of human-conducted natural language processing (Kommineni et al., 2024). Likely we will be witnessing an evolutionary change in knowledge organization and representation under the fast-developing AI era. While computer scientists focus on the pragmatic side of KO and KR by exploiting LLMs and prompt engineering to improve the accuracy, scalability, and depth of knowledge captured, the values in theoretical advancements should not be overlooked. The lack of communication between AI and KOS communities may hinder the applicability of ontologies as knowledge resources. With the domain shifting towards generative AI, more work is necessary to refine the ‘core’ and paradigm of this interdisciplinary field. This may seem trivial to application-driven science, but a ‘*step back*’ from the actual phenomena is to find a broader characterization that encompasses the instances at hand (Lyytinen et al., 2004). The paradigmatic similarities in KR between KO and AI offer not only theory foundations but also practicalities for KO to contribute its unique value for knowledge representation (Qin, 2020).

Fundamentally, KOSs must be prepared to handle various types of knowledge generated by AI models. Measurements for semantic similarity should be updated to comply with new relations between entities and relationships as new terminologies, or even new knowledge domains, emerge. Guidelines for misinformation detection are necessary if ontologies are using LLM-generated knowledge. Questions on the trustworthiness of AI can also impact the reliability of ontologies (Kaur et al., 2023), especially when the construction of LLMs and algorithms are mostly hidden in ‘black boxes’. In terms of medical and health data, the fairness of AI using knowledgebase data can cause ethical issues. Patient privacy and confidentiality are necessary factors that decide whether we are entrusted to use this data in ML models and not be shared for all other purposes.

Applications like disease-gene association identification is closely intertwined with clinical decision-making, which can have a direct impact on medical practice and communication between physicians and patients (Lötsch et al., 2022). A new ethical framework may be in order to balance the need of society and future patients with legitimate expectations of privacy (Haendel et al., 2018), especially with the involvement of AI models.

Conclusion

This paper focuses on using GO as case for training ML models to serve a type of computational biological research – disease-gene association identification. We calculated the functional similarity of gene products represented by the GO annotation semantic similarity, and trained three supervised ML models – SVM, RF, and GB. From experimenting on this workflow, we evaluate the role of ontologies as knowledgebases for large data biomedical research. Applying theories in knowledge organization and representation, we argue that knowledge organization and computer science domains require more communication and synchronization in the face of emerging AI and LLM technologies in order to accommodate to AI-generated knowledge and policies. We conclude that ontologies have played a crucial role in the discovery of biomedical knowledge and clinical decision-making by providing meaningful, structured, and reliable data.

Acknowledgements

We thank the comments provided by the anonymous reviewers for this paper.

About the authors

Qiaoyi Liu is a Ph.D. student in Information Science and Technology at Syracuse University. Her research interests are in knowledge organization (KO) and science of science (SoS). Especially, she studies biological knowledge representation and construction of ontologies guided by classification theories and semantic measurements. She is interested in knowledgebases exploited by ML models and trustworthy LLMs to generate knowledge for bioinformatics and computational biology research. She can be contacted at qliu11@syr.edu

Jian Qin is Professor of the iSchool at Syracuse University. She conducts research in metadata, knowledge modelling and representation, ontologies, research collaboration networks, research impact assessment, and data curation. Jian Qin directs a Metadata Lab, a research group focusing on big metadata analytics and knowledge modelling. Her research has received funding from US NSF, NIH, IMLS, among others. She publishes widely with more than 100 journal and conference papers in the field of information science, scientometrics, knowledge organization, and metadata and been invited to give keynotes, lectures, and presentations at conferences and institutions inside and outside of the U.S. She is the co-author of the book Metadata and co-editor for several special journal issues on knowledge discovery in databases and knowledge representation. She received the 2020 Frederick G. Kilgour Award for Research in Library and Information Technology. Jian Qin holds a Ph.D. from University of Illinois at Urbana-Champaign. She can be contacted at jqin@syr.edu

References

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene Ontology: Tool for the unification of biology. Nature Genetics, 25(1), 25–29. https://doi.org/10.1038/75556

Asif, M., Martiniano, H. F. M. C. M., Vicente, A. M., & Couto, F. M. (2018). Identifying disease genes using machine learning and gene functional similarities, assessed through Gene Ontology. PLOS ONE, 13(12), e0208626. https://doi.org/10.1371/journal.pone.0208626

Ata, S. K., Wu, M., Fang, Y., Ou-Yang, L., Kwoh, C. K., & Li, X.-L. (2021). Recent advances in network-based methods for disease gene prediction. Briefings in Bioinformatics, 22(4), bbaa303. https://doi.org/10.1093/bib/bbaa303

Auslander, N., Gussow, A. B., & Koonin, E. V. (2021). Incorporating Machine Learning into Established Bioinformatics Frameworks. International Journal of Molecular Sciences, 22(6), Article 6. https://doi.org/10.3390/ijms22062903

Bello, S. M., Shimoyama, M., Mitraka, E., Laulederkind, S. J. F., Smith, C. L., Eppig, J. T., & Schriml, L. M. (2018). Disease Ontology: Improving and unifying disease annotations across species. Disease Models & Mechanisms, dmm.032839. https://doi.org/10.1242/dmm.032839

Bench-Capon, T. J. M. (1990). Knowledge representation: An approach to artificial intelligence. Academic Press.

Binder, J., Ursu, O., Bologa, C., Jiang, S., Maphis, N., Dadras, S., Chisholm, D., Weick, J., Myers, O., Kumar, P., Yang, J. J., Bhaskar, K., & Oprea, T. I. (2022). Machine learning prediction and tau-based screening identifies potential Alzheimer's disease genes relevant to immunity. Communications Biology, 5(1), 125. https://doi.org/10.1038/s42003-022-03068-7

Carlson, M., Falcon, S., Pages, H., & Li, N. (2019). Org. Hs. Eg. Db: Genome wide annotation for Human. R Package Version, 3(2), 3.

Chang, J., Wang, S., Ling, C., Qin, Z., & Zhao, L. (2024). Gene-associated Disease Discovery Powered by Large Language Models (No. arXiv:2401.09490). arXiv. https://doi.org/10.48550/arXiv.2401.09490

Duck, G., Nenadic, G., Filannino, M., Brass, A., Robertson, D. L., & Stevens, R. (2016). A Survey of Bioinformatics Database and Software Usage through Mining the Literature. PLOS ONE, 11(6), e0157989. https://doi.org/10.1371/journal.pone.0157989

Geschwind, D. H., & State, M. W. (2015). Gene hunting in autism spectrum disorder: On the path to precision medicine. The Lancet Neurology, 14(11), 1109–1120. https://doi.org/10.1016/S1474-4422(15)00044-7

Google, Microsoft, Yahoo, & Yandex. (2024). Schema.org [Organization]. Schema.Org. https://schema.org/

Guarino, N., & Giaretta, P. (1995). Ontologies and knowledge bases: Towards a terminological clarification. In Towards Very Large Knowledge Bases. (pp. 25–32). IOS Press. https://www.loa.istc.cnr.it/old/Papers/KBKS95.pdf

Haendel, M. A., Chute, C. G., & Robinson, P. N. (2018). Classification, Ontology, and Precision Medicine. The New England Journal of Medicine, 379(15), 1452–1462. https://doi.org/10.1056/NEJMra1615014

Kim, Y. S., & Leventhal, B. L. (2015). Genetic Epidemiology and Insights into Interactive Genetic and Environmental Effects in Autism Spectrum Disorders. Biological Psychiatry, 77(1), 66–74. https://doi.org/10.1056/NEJMra1615014

Krishnan, A., Zhang, R., Yao, V., Theesfeld, C. L., Wong, A. K., Tadych, A., Volfovsky, N., Packer, A., Lash, A., & Troyanskaya, O. G. (2016). Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. Nature Neuroscience, 19(11), 1454–1462. https://doi.org/10.1038/nn.4353

Levesque, H. J., & Lakemeyer, G. (2022). The logic of knowledge bases (Second edition). College Publications.

Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. Nature Reviews Genetics, 16(6), 321–332. https://doi.org/10.1038/nrg3920

Min, S., Lee, B., & Yoon, S. (2016). Deep learning in bioinformatics. Briefings in Bioinformatics, bbw068. https://doi.org/10.1093/bib/bbw068

National Center for Biomedical Ontology. (2024). Welcome to BioPortal, the world's most comprehensive repository of biomedical ontologies. [Organization]. BioPortal. https://bioportal.bioontology.org/

Ng, A. (Director). (2024, July 4). A Chat with Andrew on MLOps: From Model-Centric to Data-Centric AI[OL]. [Video recording]. https://www.youtube.com/watch?v=06-AZXmwHjo

Pesquita, C., Faria, D., Falcão, A. O., Lord, P., & Couto, F. M. (2009). Semantic Similarity in Biomedical Ontologies. PLOS Computational Biology, 5(7), e1000443. https://doi.org/10.1371/journal.pcbi.1000443

Piñero, J., Ramírez-Anguita, J. M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., & Furlong, L. I. (2020). The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Research, 48(D1), D845–D855. https://doi.org/10.1093/nar/gkz1021

Qin, J. (2002). Evolving Paradigms of Knowledge Representation and Organization: A Comparative Study of Classification, XML/DTD, and Ontology. ADVANCES IN KNOWLEDGE ORGANIZATION, 8, 465–471.

Qin, J. (2020). Knowledge Organization and Representation under the AI Lens. Journal of Data and Information Science, 5(1), 3–17. https://doi.org/10.2478/jdis-2020-0002

Qin, J., & Liu, Q. (2024). Organizing Knowledge in Knowledgebases: A Case Study. Knowledge Organization for Resilience in Times of Crisis: Challenges and Opportunities, 393–400.

Rada, R., Mili, H., Bicknell, E., & Blettner, M. (1989). Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man, and Cybernetics, 19(1), 17–30. https://doi.org/10.1109/21.24528

Radivojac, P., Peng, K., Clark, W. T., Peters, B. J., Mohan, A., Boyle, S. M., & Mooney, S. D. (2008). An integrated approach to inferring gene–disease associations in humans. Proteins: Structure, Function, and Bioinformatics, 72(3), 1030–1037. https://doi.org/10.1002/prot.21989

Resnik, P. (1999). Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. Journal of Artificial Intelligence Research, 11, 95–130. https://doi.org/10.1613/jair.514

The Gene Consortium. (2023). The Gene Ontology knowledgebase in 2023. Gene, 224(1), 1–14. https://doi.org/10.1093/genetics/iyad031

Voineagu, I., & Eapen, V. (2013). Converging Pathways in Autism Spectrum Disorders: Interplay between Synaptic Dysfunction and Immune Responses. Frontiers in Human Neuroscience, 7. https://doi.org/10.3389/fnhum.2013.00738

Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S., & Chen, C.-F. (2007). A new method to measure the semantic similarity of GO terms. Bioinformatics, 23(10), 1274–1281. https://doi.org/10.1093/bioinformatics/btm087

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 160018. https://doi.org/10.1038/sdata.2016.18

Yang, K., Lu, K., Wu, Y., Yu, J., Liu, B., Zhao, Y., Chen, J., & Zhou, X. (2021). A network-based machine-learning framework to identify both functional modules and disease genes. Human Genetics, 140(6), 897–913. https://doi.org/10.1007/s00439-020-02253-0

Yousef, M., Sayıcı, A., & Bakir-Gungor, B. (2021). Integrating Gene Ontology Based Grouping and Ranking into the Machine Learning Algorithm for Gene Expression Data Analysis. In G. Kotsis, A. M. Tjoa, I. Khalil, B. Moser, A. Mashkoor, J. Sametinger, A. Fensel, J. Martinez-Gil, L. Fischer, G. Czech, F. Sobieczky, & S. Khan (Eds.), Database and Expert Systems Applications—DEXA 2021 Workshops (pp. 205–214). Springer International Publishing. https://doi.org/10.1007/978-3-030-87101-7_20

Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y., & Wang, S. (2010). GOSemSim: An R package for measuring semantic similarity among GO terms and gene products. Bioinformatics, 26(7), 976–978. https://doi.org/10.1093/bioinformatics/btq064

Zeng, M. L. (2008). Knowledge Organization Systems (KOS). KNOWLEDGE ORGANIZATION, 35(2–3), 160–182. https://doi.org/10.5771/0943-7444-2008-2-3-160