Technological Forecasting & Social Change 164 (2021) 120513

Contents lists available at ScienceDirect

Technological Forecasting & Social Change

journal homepage: www.elsevier.com/locate/techfore

Exploring the genetic basis of diseases through a heterogeneous bibliometric network: A methodology and case study

Mengjia Wu, Yi Zhang *, Guangquan Zhang, Jie Lu

Australian Artificial Intelligence Institute, Faculty of Engineering and Information Technology, University of Technology Sydney, Australia

ARTICLE INFO

Keywords:
Bibliometrics
Network analytics
Disease genetic basis
Word embedding

ABSTRACT

Literature-based knowledge (LBD) discovery is a practical approach to inferring the associations between diseases and genetic factors from unstructured biomedical data, i.e., the literature. However, most of the contemporary LBD methods are designed for specific cases and rely heavily on prior knowledge. In this paper, we propose an adaptable and transferable methodology that not only summarizes the genetic factors known to be associated with a queried disease but also predicts likely associations that have yet to be identified. The framework incorporates different biomedical entities in a heterogeneous co-occurrence network. Three centrality indicators, coupled with a novel measure based on intersection ratios, capture the importance and specificity of each factor to the disease under study. Undiscovered, but likely, associations are identified through a semantic similarity matrix generated by our Bioentity2Vec model and an innovative weighted link prediction algorithm. The final outputs are ranked lists of the most relevant known or potential biomedical associations. To both test and showcase the methodology, we conducted a case study on atrial fibrillation. The analysis yields specific insights into the key biomedical entities associated with this disease. Moreover, it demonstrates the kind of valuable decision support this framework can provide to medical researchers, policymakers and public health administrations.

1. Introduction

In modern medicine, deciphering the genetic basis of disease plays a vital role in its diagnosis, treatment, and prevention. However, for most disorders and abnormalities, it is not yet known whether genes, gene mutations, genetic variations, etc. play a pathogenetic role (Cookson et al., 2009; Goldstein, 2009). The high cost of genetic linkage analysis (Ott, 1999) and genome-wide association studies (GWAS) (Bush and Moore, 2012) has spawned an urgent need to prioritize candidate factors for analysis. In past decades, researchers have established medical ontologies and curated molecular networks to analyze and infer molecular interactions for diseases based on accumulated experimental and clinical experience (Barabási et al., 2011). Although these curated knowledge bases provide a source of structured data for genetic discoveries, their use is still limited for a number of reasons, including: 1) the limited entity categories and the restrictions from the fixed knowledge base framework; 2) the time lag between a discovery and the dataset being updated; and 3) the enormous cost of establishing and maintaining these knowledge bases.

However, advanced text mining techniques combined with a fast-growing body of rich biomedical texts may provide an accessible, real-time, and economically-viable pathway to solving those issues (Opap and Mulder, 2017) through literature-based knowledge discovery. Techniques such as co-occurrence analysis (Cohen et al., 2005), meta-analysis (Wang et al., 2017), centrality measurement (Al-Aamri et al., 2019), text mining (Mallory et al., 2016), and machine learning (Kim et al., 2017) have broadly empowered literature searches as an efficient way of exploring the genetic basis of various diseases, but the contemporary methods have some significant shortcomings. The main problem is that most are designed with a singular focus on a specific disease. There are very few generalized models. Further, there is a tendency toward the imbalance between quantitative approaches and expert knowledge (Al-Aamri et al., 2019). Too often, getting good results relies on substantial human intervention and prior knowledge that can be hard to access. These constraints are particularly problematic for rare diseases and diseases where multiple genes may contribute to a condition.

Aiming to address these concerns, we propose a general and

* Corresponding author.
E-mail addresses: mengjia.wu@student.uts.edu.au (M. Wu), yi.zhang@uts.edu.au (Y. Zhang), Guangquan.zhang@uts.edu.au (G. Zhang), jie.lu@uts.edu.au (J. Lu).

https://doi.org/10.1016/j.techfore.2020.120513
Received 30 June 2020; Received in revised form 22 October 2020; Accepted 3 December 2020
Available online 16 December 2020

adaptable bibliometric methodology for investigating the genetic factors associated with a specific disease. To provide a thoroughly comprehensive analysis, four types of biomedical entities are considered: diseases, chemicals, genes, and genetic variations. The methods are data-driven and do not require human intervention, which guarantees adaptability to different cases. Further, the framework incorporates semantic similarity and weighted link prediction techniques to infer likely associations that have yet to be identified.

The main components include a heterogeneous bibliometric network, a Bioentity2Vec model, a suite of network analytics indicators, and a link prediction algorithm. The nodes of the bibliometric network represent the four types of biomedical entities, and the edges represent sentence-level co-occurrence between the nodes. The Bioentity2Vec model follows the same approach as Word2Vec (Mikolov et al., 2013), where all the entities are represented as computable vectors and then used to generate an adjacency matrix of pairwise semantic similarity. The network indicators include a series of centrality measurements that measure the importance of each entity, plus a novel indicator called *intersection ratio* that measures the specificity of an entity to the disease under study. The link prediction algorithm is a modified version of the resource allocation algorithm to incorporate semantic similarity. Its purpose is to identify likely, but as yet undiscovered, genetic associations.

To validate our processes and to demonstrate how the methodology works, we conducted a case study on a corpus of 54,219 academic papers related to atrial fibrillation (AF). In a comparison test with a divided dataset designed to provide ground truths, our link prediction method identified 74% of the factor associations that would come to emerge spanning genes and genetic variants. In the same test on data up to 2020, we discovered strong evidence for five potential undiscovered associations and mediocre evidence for another five.

There are several novel aspects of our work. First, the literature-based method in our work that does not rely on prior biomedical knowledge. Further, our methodology the heterogeneous networks in the analysis for biomedical entity interactions. Lastly, the combination of contextual semantic similarity and topological similarity enhanced with link prediction is a powerful new technique that has broad applications in information science.

The rest of this paper is organized as follows. In Section 2, we briefly review relevant work on network medicine and literature-based discovery as it pertains to exploring the genetic basis for diseases. Section 3 presents the details of our methodology. The case study on AF appears in Section 4, along with the results. Section 5 wraps up the study with a discussion and the conclusions.

2. Related work

This section provides a critical review of previous studies on network medicine and literature-based knowledge discovery with a particular focus on the data-driven exploration of the genetic factor-disease associations.

2.1. Network medicine

As the name suggests, the network medicine methodology principally involves constructing a network of biomedical entities and using that network to analyze the interactions between them (Barabási et al., 2011). With the accumulation of experimental findings and clinical evidence, many biomedical datasets and interaction networks have been created and curated, including protein-protein interaction (PPI) networks (Szklarczyk et al., 2019), metabolic networks (Lawson et al., 2017; Schellenberger et al., 2010), regulatory networks (Clemente-Casares et al., 2016; Newburger and Bulyk, 2009), RNA networks (Anastasiadou et al., 2018), gene co-expression networks (van Dam et al., 2018), and so on.

Further, a substantial amount of research has been undertaken to unlock the knowledge within these constructs using different network analytics approaches. For example, Lei and Ruan (2013) proposed a topological similarity-based approach to reduce the sparsity of a protein-protein interaction network, reconstructing a more condensed network for genetic analysis that has better computational efficiency and more accurate predictions. Ganegoda et al. (2014) developed a method for constructing tissue-specific gene networks from a whole disease-gene network and applied a path-based similarity measurement to validate its usefulness. Valdeolivas et al. (2019) constructed a heterogeneous network containing diseases, genes, and proteins as entities and, further, implemented a random walk on the network to infer disease-gene interactions. However, despite all these successful explorations of fundamental knowledge bases, only a narrow slice of the possible biomedical entities is covered in each study. Plus, the results do not include very recent discoveries, and the economic cost and human effort to establish and continue maintaining the datasets that drive these solutions is enormous.

2.2. Literature-Based discovery for disease genetic basis

As an alternative to network analytics, literature-based knowledge discovery provides a more widely accessible pathway to exploring the genetic basis for disease (Zhang et al., 2018b, 2016). One of the earliest attempts to discover genetic knowledge from the literature, by Stapley and Benoit (2000), was to extract terms from the articles based on a dictionary and use those terms in conjunction with a set of rules to construct a gene co-occurrence network. Jenssen et al. (2001) further validated the usefulness of co-occurring patterns by visualizing a global human genome network comprising millions of articles from three large-scale datasets of biomedical literature. The resulting network revealed many meaningful associations between co-occurring gene names at the document-level. To link genes with diseases through text analysis, Adamic et al. (2002) applied statistical analysis to disease-gene co-occurrences using the binomial distribution. They scored the relevance of genes and diseases to discover novel associations, shedding light on multiple biomedical entity analysis. However, these studies suffered one common hindrance, which was a high gene name recognition. In too many cases, confusing abbreviations (aliases) and the difficulty of disambiguating symbols confounded the models.

Natural language processing techniques have further improved the efficiency and accuracy of biomedical entity extraction (Habibi et al., 2017; Mallory et al., 2016; Pletscher-Frankild et al., 2015; Wei et al., 2013). Garten et al. (2010) employed a text mining-based extractor to perform sentence-level drug and gene co-occurrence analysis. The results show the superiority of using a text-derived network over a manually-curated network of drug-gene relationships to make predictions (Garten et al., 2010). Özgür et al. (2008) established a disease-specific gene interaction network and used network centralities to infer genes with potential links to prostate cancer along with already-known seed genes. Al-Aamri et al. (2019) developed an approach based on network centrality, where the classifier is trained using a bootstrapping method. As a result, the model was able to parse the entire human genome. Some studies also leverage the semantic similarity between entities. For example, Coulet et al. (2010) built semantic networks of pharmacogenomic entities based on text data and inferred the interactions between them. Schlicker et al. (2010) improved gene prioritization accuracy by involving the semantic similarities generated from an ontology of genetic terms. With all the indicators in a basket, Heo et al. (2019) combined entity co-occurrence with word embedding techniques to produce a comprehensive index with which to measure the relationships between entities related to Alzheimer's disease.

The framework presented in this paper is an extension of previous work by us. In a 2018 study (Zhang et al., 2018a), we developed a bibliometric technique based on Word2Vec (Mikolov et al., 2013) to improve the accuracy of semantic similarity measurements with textual

data. In a subsequent study, published in *Intelligent Bibliometrics* (Zhang et al., 2020a), we piloted some intelligent models for recognizing bibliometric patterns. This work extends those models to the biomedical arena in a generalized application designed to explore the literature for this known and/or likely genetic basis of any given disease.

3. Methodology

The framework of the proposed method is illustrated in Fig. 1. The five blocks of the methodology include entity extraction, heterogeneous network construction, Bioentity2Vec training, core entity identification, and semantic-enhanced link prediction.

3.1. Entity extraction and heterogeneous network construction

As mentioned, the framework considers four types of biomedical entities: diseases, chemicals, genes, and genetic variations. 1) Diseases include disorders, symptoms, risk factors, and complications. 2) Chemicals covers chemical elements, clinical medications, and other compounds. 3) Genes are defined as the basic unit of heredity, occupying a fixed position on the chromosome. 4) Genetic variants include DNA mutations (i.e., a permanent change in a DNA sequence), protein mutations (i.e., proteins encoded with a mutated gene) and single nucleotide polymorphisms (SNPs) (i.e., normal variations of a single nucleotide in a gene sequence) (Arias et al., 1991).

These four entities are represented as nodes in a weighted heterogeneous network. Working under the hypothesis that sentence-level co-occurrence indicates a stronger association between pairwise entities than document-level co-occurrence, the edges reflect co-occurrence frequency at sentence-level. The weights are derived from an adjacency matrix A, in which V_i^m is the mth node in the ith category and

$$A_{V_i^m V_j^n} = \begin{cases} CF\left(V_i^m, V_j^n\right) & \left(if\ V_i^m\ and\ V_j^n\ co-occur\ in\ a\ sentence\right) \\ 0\# \end{cases}.$$

$CF_{V_i^m V_j^n}$ is the sentence co-occurrence frequency between V_i^m and V_j^n.

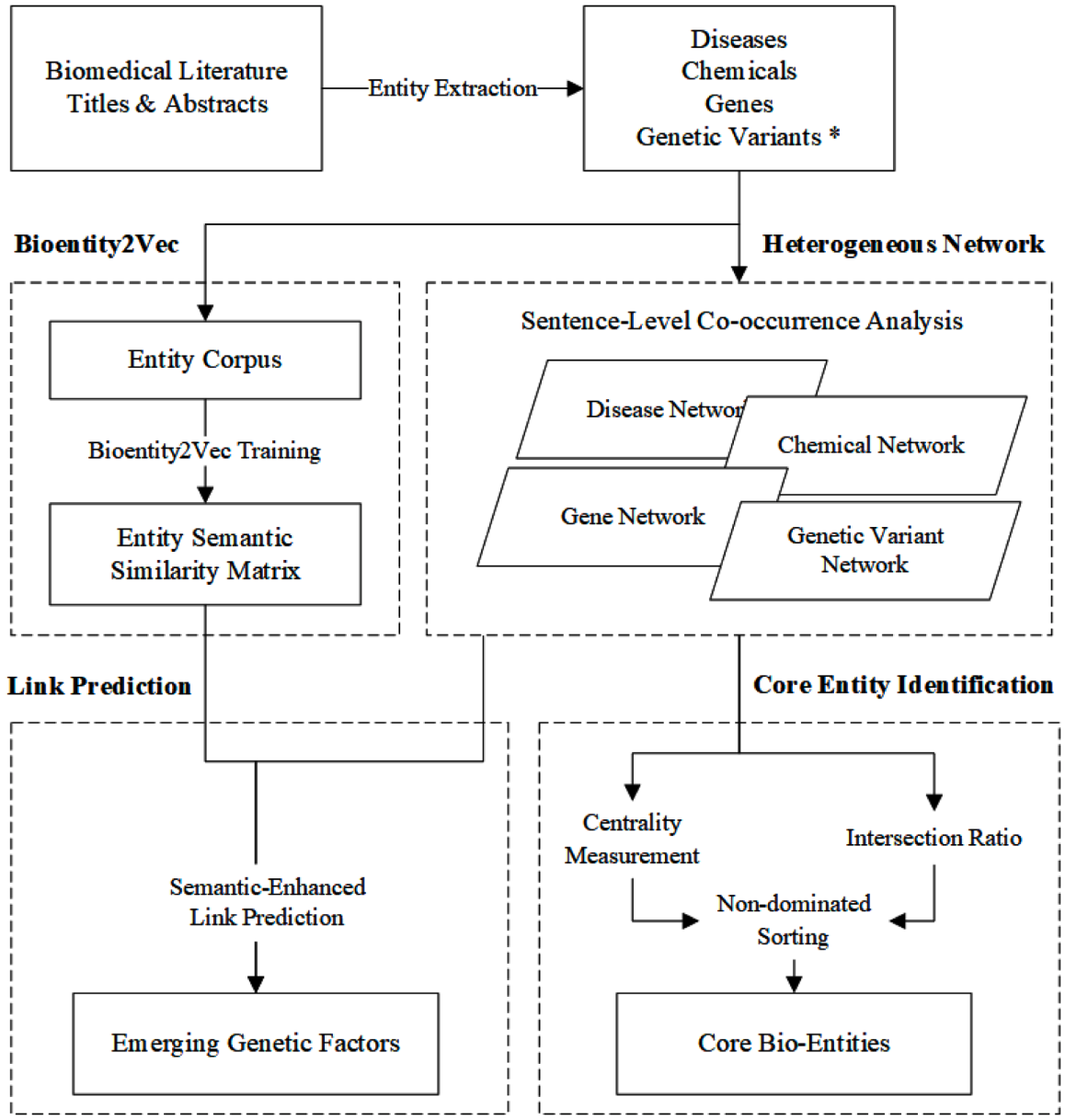

Fig. 1. The framework of our proposed methodology
*Note: genetic variants include DNA mutations, protein mutations and single nucleotide polymorphisms (SNP).

This network can also be represented as a graph:

$$G = \left(V_K,\ E_{K(K+1)/2}\right)$$

where V is the set of K entity categories and E is the set of $K(K+1)/2$ types of edges connecting the different categories of nodes. An illustration of the network is provided in Fig. 2.

3.2. Bioentity2Vec modelling

Sparked by the idea of the well-regarded Word2Vec natural language model (Mikolov et al., 2013), our semantic similarity measures are taken from a context-based perspective using a model we developed called Bioentity2Vec. Like Word2Vec, Bioentity2Vec converts biomedical entities into vectors by projecting one-hot representations into a lower dimension while largely preserving the semantic meaning of the content. In our case, the biomedical entities are treated as words, and those words placed in sequence constitutes the training corpus. We selected Skip-Gram as our training algorithm since it offers a better fit on small datasets (Mikolov et al., 2013). A summary of the Skip-Gram training process follows.

Given an entity $E(i)$ in a corpus, the probabilities of other entities in a certain window size w are predicted based on the given central entity $E(i)$ (Rong, 2014). The global objective is to maximize the average conditional probability for all windows in the corpus, which is formulated as:

$$LF = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{-w\le j\le w,\ i\neq 0}\log_2 P(E(i+j)E(i))\right)$$

The first step is to calculate the pairwise similarity of entities via cosine similarity and then generate an adjacency semantic similarity matrix $S_{V_i^m V_j^n}$:

$$S_{V_i^m V_j^n} = \cos\left(v_{V_i^m}, v_{V_j^n}\right) = \frac{v_{V_i^m}\cdot v_{V_j^n}}{\sqrt{v_{V_i^m}\cdot v_{V_i^m}}\cdot\sqrt{v_{V_j^n}\cdot v_{V_j^n}}}$$

where $v_{V_i^m}$ is the corresponding vector of entity node V_i^m.

Applying this formula to all entity pairs produces a pairwise adjacency matrix $S_{V_i^m V_j^n}$ of semantic similarity for all entities.

3.3. Network analytics

3.3.1. Centralities

Centrality comprises three indicators, degree centrality, closeness centrality, and betweenness centrality (Freeman et al., 1979; Zhang et al., 2020c), which respectively reflect the node's capacity to aggregate, disseminate, and transfer information across a network. All three have been proven efficient in revealing key nodes in biomedical networks (Al-Aamri et al., 2019). Formal definitions follow.

Degree Centrality (DC): This indicator measures the direct influence of a node on other nodes by calculating the proportion of its degree. An entity with a high degree centrality indicates that it has direct interactions with many other entities. It is calculated as:

$$DC\left(V_i^m\right) = \frac{\sum_{j=1}^{K}\sum_{n=1}^{|V_j|}A_{V_i^m V_j^n}}{|V_K|-1}$$

where $|V_K|$ is the number categories K of nodes in the network and $|V_j|$ is the node number in the jth category;

Closeness Centrality (CC): This indicates a node's topological distance to all other nodes in the network, reflecting the global impact of a node towards all other nodes within the network. It is calculated as follows:

$$CC\left(V_i^m\right) = \frac{|V_K|-1}{\sum_{j=1}^{K}\sum_{n=1}^{|V_j|}d_{V_i^m V_j^n}}$$

where $d_{V_i^m V_j^n}$ is the topological distance from node V_i^m to node V_j^n;

Betweenness Centrality (BC): This indicator measures a node's capability of connecting any other two nodes. In a network, a high betweenness centrality indicates that the node has a strong potential to be a crucial connector or transmitter. It is calculated by the sum of possibilities that any shortest paths connecting two other nodes go through the target node:

$$BC\left(V_i^m\right) = \frac{2\sum_{x,\ y=1}^{K}\sum_{a=1}^{|V_x|}\sum_{b=1}^{|V_y|}\frac{\sigma\left(V_x^a V_y^b\right)_{V_i^m}}{\sigma\left(V_x^a V_y^b\right)}}{(|V_K|-1)(|V_K|-2)}\quad\left(V_i^m \neq V_x^a \neq V_y^b\right)$$

where $\sigma(V_x^a V_y^b)$ is the number of all shortest paths from node V_x^a to V_y^b and $\sigma(V_x^a V_y^b)_{V_i^m}$ is the number of these paths that pass through node V_i^m.

3.3.2. Intersection ratio

Intersection ratio is a novel indicator we designed to distinguish entities specifically associated to the target disease. While all centrality indicators reflect some aspect of a node's significance in the global network, some entities with high centralities may not be associated with the target disease at a particularly high level. Those entities are usually general terms representing fundamental chemicals or genetic factors related to a relatively broad range of conditions. Thus, we need to distinguish the general entities from those entities that are highly relevant to the target disease. To this end, we developed an indicator we call *intersection ratio,* which is based on a Jaccard coefficient (Niwattanakul et al., 2013). This indicator reflects an entity's specificity as the rate of a node's interaction with the target disease over all other diseases:

$$IR\left(V_i^m\right) = \frac{w\left(V_i^m, V_{disease}^t\right)}{\sum_{a=1}^{|V_{disease}|}w\left(V_i^m, V_{disease}^a\right)}$$

where $V_{disease}^t$ represents the node of the target disease, and $w(V_i^m, V_{disease}^t)$ refers to the weight of the edge connecting V_i^m and $V_{disease}^t$.

Traditionally, bibliometrics-based indicators are either combined by certain strategies (e.g., entropy) into a unique value or are pairwise visualized based on diverse actual requirements (Zhang et al., 2017). However, our aim is to build a general framework; therefore, we introduced the non-dominated sorting algorithm to rank the entities based on a combination of the four metrics. Technically, non-dominated sorting is a multi-objective optimization procedure that compares samples containing multiple objectives or dimensions and ranks them according to their "dominance" over each other (Yuan et al., 2014). An entity A would dominate an entity B if A was better than B according to at least one of the four indicators, but was no worse than B in any of the others. Once sorted, the items are divided into several consecutive Pareto fronts according to their domination counts. For example, if entity A is better than entity B in all four measures, it will be assigned to the dominant Pareto front. With top ranks in all metrics, the entities in this front are deemed the ones with the strongest associations to the disease under study.

The pseudo-code for the non-dominated sorting algorithm is shown in Fig. 3.

The output of this set of measurements is four different lists of entities related to the target disease, i.e., core diseases, chemicals, genes, and genetic variations, ranked in non-dominated order.

3.3.3. Semantic similarity-enhanced link prediction

Link prediction describes approaches that estimate the probability of particular links emerging in a network in the future (Liben-Nowell and Kleinberg, 2007). The results from our pilot study show that, of all the

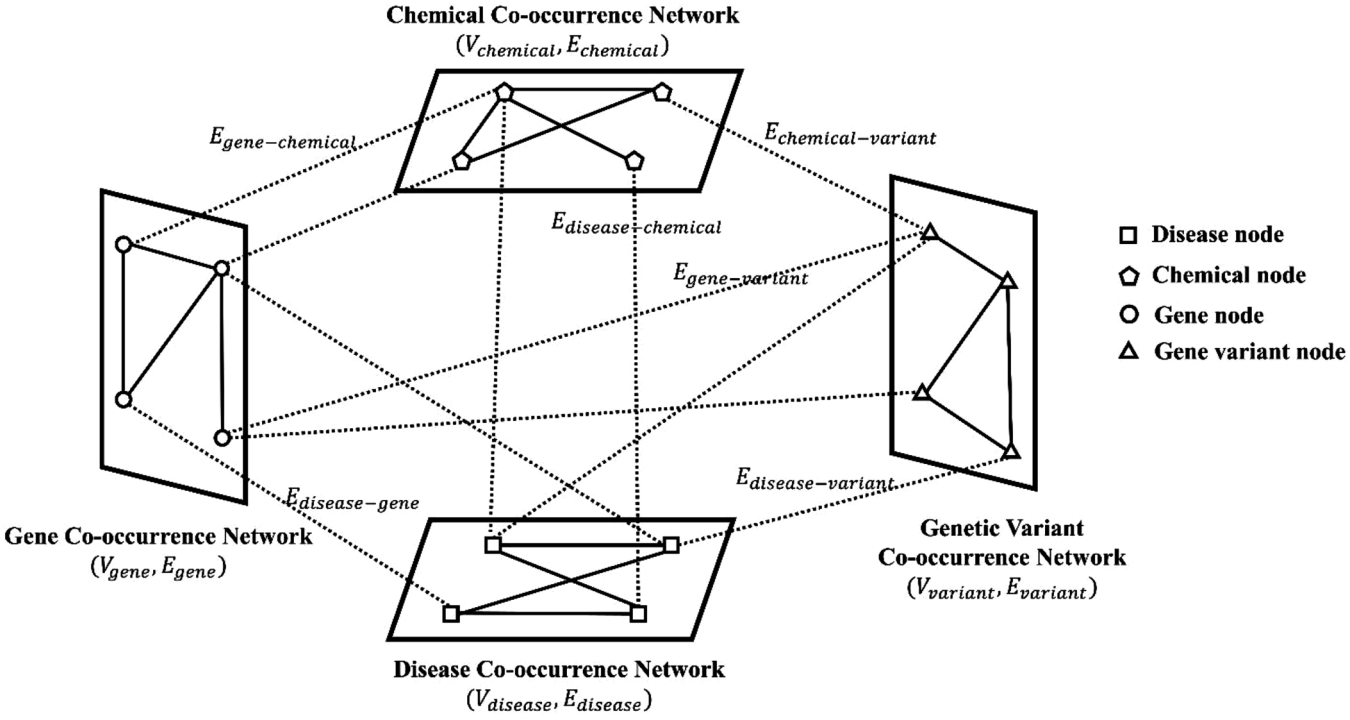

Fig. 2. The types of co-occurrence associations in the network.

for V_i in V_K:

  for node V_i^m in V_i:

    $Domination[V_i^m] = 0$ # Initialize the Domination Counts

    for node V_i^n in category i $(m \neq n)$:

      if $M_1(V_i^m) \geq M_1(V_i^n)$ *and* $M_2(V_i^m) \geq M_2(V_i^n)$ *and* ... $M_d(V_i^m) \geq M_d(V_i^n)$ *and*

        $(M_1(V_i^m) = M_1(V_i^n)$ *and* $M_2(V_i^m) = M_2(V_i^n)$ *and* ... $M_d(V_i^m) = M_d(V_i^n)) == False$:

      # $M_d(V_i^m)$ *refers to the dth dimensional measurement of* V_i^m

      $Domination[V_i^m]$ += 1

Fig. 3. The pseudo-code of the non-dominated sorting algorithm

neighbor-based comparisons methods, resource allocation (RA) (Zhou et al., 2009) is the most accurate (Zhang et al., 2020b; Zhang et al., 2020c). The original RA algorithm follows the assumption that every node in a network has one unit of a resource, and a common neighbor to two nodes will act as a transmitter, evenly distributing its resource to the connected nodes. The RA index of an unconnected pair of nodes is the sum of all resources obtained from all the neighbors common to the two nodes. In simple terms, it reflects the potential for a direct link to form between the nodes. The higher the value, the greater the possibility is for a future link.

Inspired by Lü and Zhou (2010), who developed a weighted version of this algorithm, we conjecture that assessing the semantic similarity between two nodes and using that to weight to the RA index will increase the accuracy of the link prediction. Hence, we incorporated an additional procedure into the algorithm that involves the semantic matrix of biomedical entities generated by the Bioentity2Vec model. Thus, the final RA index is calculated as:

$$MRA_{V_i^m V_j^n} = S_{V_i^m V_j^n} \cdot \sum_{V^t \in \Gamma(V_i^m) \cap \Gamma(V_j^n)} \frac{CF(V_i^m, V^t)\left|S_{V_i^m V^t}\right| + CF(V^t, V_j^n)\left|S_{V^t V_j^n}\right|}{\sum_{V^k \in \Gamma(V^t)} CF(V^k, V^t) S(S_{V^k, V^t})}$$

where V_i^m is the target disease, and V_j^n belongs to the set of genetic factors that have never before co-occurred with the target disease.

Applying the modified link prediction approach in a pairwise manner (V_i^m, V_j^n) generates the final output, which is a ranked list of genetic factors and their corresponding RA index scores. The assumption underlying the prediction that a genetic factor is associated with a disease is: if the target disease node V_i^m and the genetic factor node V_j^n do not co-occur but they share at least one common neighbor, then they have the potential to be directly associated. The RA score tells us how strong that potential is. The common neighbor could be any one of the four entities –

for example, they may both be associated with another genetic factor, or they may both be reactive to the same chemical.

4. Case study

Atrial fibrillation (AF) is one of the most common forms of cardiac arrhythmia. Disease progress is closely related to atrial size and the extent of atrial fibrosis, both of which are affected by genetic factors. Although several gene groups and genetic mutations have been linked to AF, clinical evidence and mechanistic explanations are still far from sufficient to begin integrating our knowledge of these genetic risk factors into clinical practice (Feghaly et al., 2018). For these reasons, exploring the associations between genes and AF as our case study not only serves as an assessment of the proposed method, it may also have practical significance for advancing research frontiers in the AF area.

4.1. Data collection

PubMed is a search engine for biomedical literature, comprising more than 30 million citations across the MEDLINE database, life science journals, and other online book resources. Additionally, it provides various toolkits for analyzing the literature. We used the term "atrial fibrillation" with a MeSH search strategy limited to the "species" human across PubMed titles to guarantee precise AF-related search results. No restrictions were placed on the publication date. In all, 54,219 records were retrieved from the following searching query:

"("Atrial Fibrillation"[Mesh] AND Humans[Mesh])" Search Date: 28 April 2020

4.2. Entity extraction and network establishment

As mentioned, high error rates are a common challenge with tasks involving gene name recognition. To lessen this problem, we assembled the extractor's vocabulary list by combining terms from three different biomedical dictionaries:

- Medical Subject Headings (MeSH)[1] is a medical thesaurus, provided by PubMed, that contains the normal standardized concepts of diseases and chemicals.
- NCBI *Homo-Sapiens* Gene Dictionary,[2] provided by the United States National Institute of Health (NIH), covers the known genes of *Homo sapiens* species.
- dbSNP database[3] is a register of known sequence variants in the human genome, established in 1999. It contains the discovered DNA mutations, protein mutations and SNPs, each record is associated with an identical SNP ID.

We selected Pubtator[4] as the extractor. Pubtator is a deep learning-based tool for biomedical entity extraction, which was developed by the National Library of Medicine (NLM) (Wei et al., 2019). It can automatically extract categorized biomedical concepts from the titles and abstracts of PubMed articles.

The raw extraction process resulted in 577,809 biomedical concepts with accompanying text locations and unique identifiers. The concepts included diseases, chemicals, genes, DNA and protein mutations, SNPs, and species. We excluded the species concepts since our focus is on humans, and restricted the genetic factors to the scope of the human genome using dbSNP. We then mapped every concept back to its corresponding dictionary, removing noisy concepts (see Step 1, Table 1) and consolidating all synonyms (Step 2). After these two steps, 6318 unique biomedical entities remained. We further excluded 480 concepts that did no co-occur with any other concept (i.e., isolated nodes) to result in a final set of 5838 entities. The stepwise pre-processing tallies are given in Table 1.

Bioentity2Vec revealed 48,988 edges reflecting sentence-level co-occurrence across the 5838 nodes of the network. Among the four types of entities, there can be ten types of edges; their counts are provided in Table 2.

4.3. Identifying the core entities for af

Core entities (i.e., highly-relevant entities) have high degree, closeness and betweenness centralities, plus a high intersection ratio. We calculated these metrics for all 5838 entities. A summary of the pertinent statistics by entity type is provided in Table 3.

4.3.1. Core genes

Following the steps of the framework, we began with the genes only, applying the non-dominated sorting algorithm to the three centralities and normalizing the domination counts to reflect their global importance. We then juxtaposed this against normalized intersection ratios, which reflect the gene's specificity to AF. This produced a 2-D gene scatter map of importance vs. specificity, as shown in Fig. 4. Global importance is plotted on the X-axis; specificity is plotted on the Y-axis.

The genes of most concern to us are those in the top right corner – i.e., the genes with both high centrality dominations and a high intersection ratio, which means they are not only important but also specific to AF.

However, part of the purpose of this case study is to evaluate this framework. Hence, we corroborated these results with a manual review of three biomedical knowledge bases: Online Mendelian Inheritance in Man (OMIM),[5] the Kyoto Encyclopedia of Genes and Genomes (KEGG),[6] and the Genetics Home Reference-NIH.[7] Throughout the investigation, we divided the core genes into two loci: seed genes – genes with known functions associated with the incidence of AF; and suspected correlated genes – genes with unknown functions that are possibly related to AF but yet to be explored):

1) Seed genes (in black boxes): Nine genes with direct associations to AF are documented in knowledge bases. According to OMIM, most of the different subtypes of AF[8] are caused by mutations or variations in these nine genes. The noted gene and subtype correlations are as follows: *KCNQ1-AF subtype 3, KCNE2-AF subtype 4, KCNA5-AF subtype 7, KCNJ2-AF subtype 9, NPPA-AF subtype 6, GJA5-AF subtype 11*, and *SCN3B-AF subtype 17*. Although OMIM does not explicitly state associations with particular AF subtypes for the other two genes in this group, *KCNH2* and *NKX2–5*, the Genetic Home Reference-NIH lists them as significant genes in AF's progression. In examining the significance of those genes from literature, we found that Sinner et al. (2008) identified a positive correlation between mutation *K897T* in *KCNH2* and higher incidence of AF. Similarly, Xie et al. (2013) associated the *NKX2–5* loss-of-function mutations *p.N19D* and *p.F186S* with AF via a cohort study on 136 patients with idiopathic atrial fibrillation. It worth highlighting that our framework did place all seed genes prominently; 36%, i.e., *SCN5A, MYL4, SCN1B, SCN2B,* had either lower IR or centrality dominations, pushing them out of the top right corner toward the left or bottom.
2) Suspected gene loci (in white boxes): There are 14 genes that are frequently studied because their mutations or variations are

[1] https://www.ncbi.nlm.nih.gov/mesh/
[2] ftp://ftp.ncbi.nih.gov/gene/DATA/GENE_INFO/Mammalia
[3] https://www.ncbi.nlm.nih.gov/snp/
[4] https://www.ncbi.nlm.nih.gov/research/pubtator/api.html
[5] More information could be found at https://www.omim.org/
[6] More information could be found at https://www.genome.jp/kegg/
[7] More information could be found at https://ghr.nlm.nih.gov/
[8] More information could be found at https://omim.org/entry/608583

Table 1
Stepwise results of the pre-processing procedure.

	Raw	Step 1	Cleaned	Step 2	Cons.	No co-occ.	Nodes
Disease	440,610	Removed noisy concepts like "cardioembolic", "JAGS", "nonvitamin", etc. that could not	434,198	MeSH	2239	−199	2040
Chemical	104,072	be mapped to MeSH	101,512		2187	−183	2004
Gene	31,209	Excluded genes that do not belong to *Homo-sapiens*	26,948	NCBI Gene	1506	−93	1413
Gene variant							
- DNA mutation	223	Removed variants with unclear loci (i.e., could not be mapped to an SNP ID)	161	dbSNP	386	−5	381
- Protein mutation	770		555				
- SNP	925	–	217				
Total	**577,809**	**–**	**563,235**	**–**	**6318**	**−180**	**5838**

Table 2
Counts of the different types of edges.

	Disease (2040)	Chemical (2004)	Gene (1413)	Genetic Variant (215)
Disease (2040)	19,181	10,977	5318	469
Chemical (2004)	10,977	5248	2463	123
Gene (1413)	5318	2463	3477	654
Genetic Variant (215)	469	123	654	495

Table 3
Centrality and intersection ratio statistics.

		Disease	Chemical	Gene	Genetic Variant
Degree Centrality	Max.	0.668	0.106	0.050	0.006
	Min.	0.0001	0.0001	0.0001	0.0001
	Avg.	0.004	0.002	0.002	0.001
	Std.	0.019	0.006	0.003	0.0008
Closeness Centrality	Max.	0.739	0.493	0.471	0.433
	Min.	0.0001	0.0001	0.0001	0.0001
	Avg.	0.397	0.381	0.382	0.378
	Std.	0.063	0.068	0.085	0.074
Betweenness Centrality	Max.	0.630	0.020	0.005	0.002
	Min.	0	0	0	0
	Avg.	0.0005	0.0001	0	0
	Std.	0.014	0.0007	0.003	0.0001
Intersection Ratio	Max.	1	1	1	1
	Min.	0	0	0	0
	Avg.	0.276	0.364	0.459	0.555
	Std.	0.280	0.363	0.377	0.443

statistically proven to be associated with AF. Hence, they are strongly suspected genetic factors but the underlying mechanisms as to why are less understood than with the seed genes. The 14 genes are *KCNE1, KCNN2, KCNN3, KCNJ5, KCND3, CAV1, SCN10A, TBX5, PITX2, ZFHX3, GJA1, HCN4, CYP11B2*, and *TRPM4*, and the literature review revealed the following mutation/variation associations: *G25V* and *G60D* in *KCNE1* (Olesen et al., 2012), *rs337711* in *KCCN2* and *rs75190942* in *KCNJ5* (Christophersen et al., 2017), *rs13376333* to *KCNN3* (Ellinor et al., 2010), *rs12044963* in *KCND3, rs11773845* in *CAV1, rs6790396* in *SCN10A, rs883079* in *TBX5, rs2129977* in *PITX2, rs2359171* in *ZFHX3, rs13191450* in *GJA1, rs74022964* in *HCN4* (Roselli et al., 2020), *T-344C* in *CYP11B2* (Li et al., 2012). Düzen et al. (2017) found that *TRM4* expression was significantly upregulated in leukocytes of non-valvular atrial fibrillation patients.

Despite some room for improvement in appropriately capturing and highlighting all seed genes, we find these results to be sufficiently reliable to proceed with the case study.

4.3.2. Other core entities

We then applied non-dominated sorting to the other three entity categories and generated the corresponding core entity lists. The top 20 diseases, chemicals, genes, and genetic variants are given in Table 4. To evaluate the quality of the sorted genetic factors, we compared our results with data from the authoritative disease-gene association discovery database DisGeNET (Piñero et al., 2016). DisGeNET integrates data from various sources, including curated knowledge bases, modeled data, inferred data, and the literature.[9] Users can also rank the associations between diseases and genetic factors according to several provided metrics. We chose the gene-disease association score (GDA) and variant-disease association score (VDA)[10] as the best comparison to our results.

In the disease category, terms in roman type are the most common physiological or pathological phenomenon relevant to the presence or treatment of AF. Terms in italics are symptoms, complications, and risk factors. Awareness of these concepts is critical to understanding the treatment of AF. One noticeable term in the list is gastroesophageal reflux, which is frequently reported in AF patients, but, judging from current research progress, any association between the two is still inconclusive (Huang et al., 2019). Further studies to supplement the literature may reveal gastroesophageal reflux has underlying significance to this research area.

In the chemicals list, terms in roman are treatments, while terms in italics are critical receptors and ion channels in the pathogenesis of AF. Caffeine and Omega-3 fatty acids are two noticeable chemicals in the list. Over years of research, the association between caffeine and AF has intriguingly been reversed from a risk factor (Curatolo and Robertson, 1983) to one with potential preventive benefits (Abdelfattah et al., 2018). The inconsistency of these results warrants further research to provide clear evidence on the issue. The same is turnabout is true of Omega-3. Once lauded as a health supplement to reduce cardiovascular disease (Abdelhamid et al., 2020), Sheikh et al. (2019) now report that Omega-3 s might actually increase the incidence of AF. The controversy has yet to be settled. These two results show that the proposed method can identify debated chemicals for further exploration.

To validate the identified genetic variants, we used ClinVar (Landrum et al., 2016), SNPedia (Cariaso and Lennon, 2012), rankings from DisGeNET, and complementary evidence from the literature. The variants in italics are variants of seed genes. We found most of the factors identified have known associations with AF, indicating either their importance or specificity to AF. One exception is rs3789678, which does not appear in DisGeNET but is noted as having a significant association with AF in the literature (Zhao et al., 2015). Additionally, there was a notable mutation *rs121912507*, which refers to the *G628* gene transfer in *KCNH2*. This SNP is not directly related to the occurrence of AF, but rather is an adenovirus-mediated transgene expression that could be used as an effective gene therapy to prevent postoperative AF.

[9] More information could be found at https://www.disgenet.org/dbinfo
[10] More information about the metrics could be found at https://www.disgenet.org/dbinfo

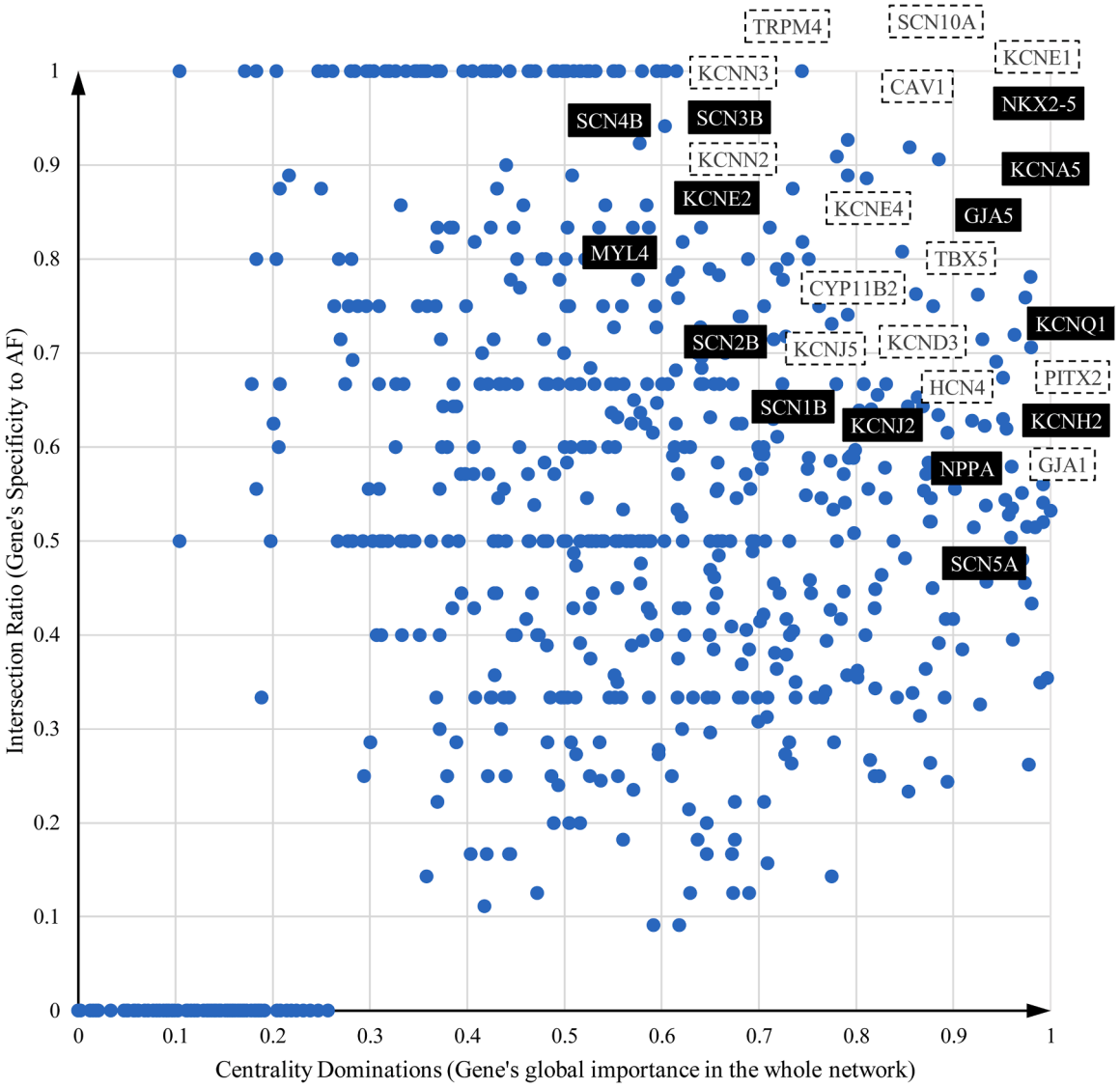

Fig. 4. Gene map plotted against the importance-specificity coordinate system

From this analysis, we are confident in concluding that our approach can identify, with relatively good accuracy, a list of biomedical entities strongly related to a given target disease. Compared to traditional approaches, such as term frequency or TF-IDF value-based sorting algorithms, this strategy produces a list of relevant, specific, and frontier entities that are not biased by the popularity of research topics.

4.4. Link prediction

Before running the link prediction algorithm, we validated its usefulness on rolled-back data. The experiment was designed as follows:

We divided the dataset into five-year brackets and constructed a network for each. k AF-linked genes or SNPs identified in the last five years were used as the true labels. We then tested our semantic-enhanced version of the RA algorithm (SERA) along with four other methods on the remaining data and compared the results. The link predictions were output as a mixed list of genes and SNPs ranked according to their RA index scores. Any gene or SNP predicted in the top n that was also in the true label set was counted as a true positive (TP), and a false negative (FN) otherwise. n is a selectable threshold that we initially set to k.

The four methods chosen for comparison were:

- RA: The original version of the resource allocation approach (Zhou et al., 2009).
- WRA: Lü & Zhou's (2010) weighted version of the resource allocation algorithm. The assumption of this algorithm is the same as RA, but the diffusing rate is measured as a weight ratio instead of as a proportion of degree centrality.
- LPI (Lü et al., 2009): Local path index is a local similarity-based index which is calculated by the weighted sum of the number of paths of length two and three. We used the default settings of 1 and 0.01 respectively for the paths of length two and three.
- RWR (Tong et al., 2008): Random walk with restart is based on a global similarity measurement for pairwise nodes. The idea of the random walk is that a particle starts from a seed node and randomly jumps to a connected node with a probability of p. The ultimate probability of the particle reaching a target node after a certain number of iterations is the possibility of forming a direct link between the seed and target node (Lovász, 1993).

To fairly evaluate the performance of all algorithms, we used the top n recall rate as the assessment metric, defined as:

$$Top\ n\ Recall = \frac{TP}{TP + FN}$$

Table 4
The top 20 biomedical entities.

Rank	Disease	Chemical	Gene Symbol	DisGeNET Ranking†	Genetic Variant SNP ID	DisGeNET Ranking
#1	*Fibrosis*	Acetylcholine	*KCNA5*	Gene: 3/73	rs2200733	AF: 1/584
#2	Atrial Remodeling	AVE0118	*GJA5*	AF: 4/939	rs13376333	AF: 5/584
#3	Atrial Flutter	*Ryanodine*	PITX2	AF: 2/939	rs2108622	Variant: 6/20
#4	Arrhythmias, Cardiac	Diltiazem	KCNE1	Gene: 10/95	rs1805127	Variant: 1/17
#5	Myocardial Stunning	Propafenone	*KCNQ1*	Gene: 6/281	rs699	Variant: 21/134
#6	*Inflammation*	Sotalol	TBX5	AF: 7/939	*rs3789678*	(Zhao et al., 2015)
#7	Tachycardia, Supraventricular	**Caffeine**	*KCNH2*	AF: 1/939	rs6795970	Variant: 5/13
#8	Atrial Premature Complexes	Isoproterenol	ZFHX3	AF: 5/939	rs3807989	AF: 4/584
#9	Mitral Valve Stenosis	Quinidine	CAV1	AF: 8/939	rs2106261	AF: 3/584
#10	*Stroke*	Verapamil	HCN4	AF: 9/939	rs10033464	AF: 2/584
#11	Heterotaxy Syndrome	*Potassium*	*NKX2–5*	AF: 12/939	rs17042171	AF: 12/584
#12	**Gastroesophageal Reflux**	*Magnesium*	CYP11B2	AF: 119/939	rs7193343	AF: 6/584
#13	Sick Sinus Syndrome	Procainamide	KCND3	Gene: 4/94	rs7164883	AF: 7/584
#14	*Stenosis, Pulmonary Vein*	Digoxin	SCN10A	AF: 10/939	rs6584555	AF: 21/584
#15	*Thromboembolism*	*Calcium*	KCNN2	Gene: 1/18	**rs121912507**	Variant: 3/4
#16	*Sleep Apnea, Obstructive*	Flecainide	KCNN3	AF: 6/939	rs1805120	Variant: 1/3
#17	*Rheumatic Diseases*	*Ibutilide*	*SCN3B*	Gene: 8/33	rs120074192	Variant: 1/10
#18	Atrioventricular Block	**Fatty Acids, Omega-3**	*NPPA*	Gene: 12/217	rs3903239	AF: 16/584
#19	Heart Disease	*Adenosine Triphosphate*	CRP	AF: 181/939	rs1152591	AF: 14/584
#20	Venous Thromboembolism	*Sodium*	*KCNJ2*	AF: 11/939	rs10824026	AF: 11/584

† "AF" refers to the gene's ranking in the list of AF-associated genes, indicating the gene's importance to AF. "Gene" refers to AF's ranking in the gene's list of associated diseases, indicating the gene's specificity to AF. The same rule applies to variant ranking.*Note: the entities in regular, italic font and bold font respectively represent different types of core entities, please refer to the following explanations for details.

The outcomes are provided in Table 5.

SERA had better recall than the other four baselines, validating its effectiveness. The top 200 predictions covered 74% of the genetic factor associations that would appear in the next five years, according to the true label set. This is a promising result, which demonstrates that our strategy can substantially reduce the heavy workload of manually seeking new candidate factors. However, the low recall for the full list of predictions (top *k*) is less than optimal. We can think of four reasons for this result: 1) Our experiment effectively simulates data streaming over time, which is a strict standard from the perspective of validation. 2) Our approach is purely data-driven without any human intervention or supervision. 3) The network is co-occurrence-based and cannot distinguish between positive and negative associations, such as "A is not associated with B". 4) The community's awareness of AF molecular mechanisms is still at a relatively early stage, and not all possible discoveries were made in the last five years. Therefore, it is reasonable to assume that some predicted associations may exist but simply require more time to uncover.

4.5. Predicting the future emerging genetic factors

The top 15 genes and SNPs from the link prediction procedure with the full network are given in Table 6. The results were validated against the DisGeNET database and the literature. Detailed explanations of the evidence found follow the table.

#1 *BGLAP* and # 10 *MGP*: These are associations to treatments for AF as opposed to AF's cause. Sato et al. (2010) and Yamagishi (2019) discovered that long-term use of Warfarin, a regular treatment used in non-rheumatic atrial fibrillation, significantly reduces the activity of *osteocalcin (BGLAP)* which protects bones from fracture and inhibits the effectiveness of matrix *Gla-protein (MGP)* in preventing vascular calcification. Those discoveries have led to alternative recommended treatments for AF patients with a high risk of fracture or vascular calcification, such as non-vitamin K oral anticoagulants (NOACs).

Table 5
Algorithm comparison (Recall).

	RWR	LPI	RA	WRA	SERA (proposed)
Top *k* Recall	0.183	0.132	0.205	0.212	**0.283**
Top 100 Recall	0.445	0.181	0.436	0.392	**0.502**
Top 200 Recall	0.621	0.576	0.714	0.632	**0.742**

#2 *rs4762:* Despite Zhao et al. (2015) finding that this mutation is not significantly associated with AF, Kuken et al. (2020) recently published a study showing that AGTT 174 M (rs4762) is associated with the occurrence of AF in the Han and Uyghur ethnic groups in Xinjiang, China. These conflicting results suggest that this SNP may perform differently for different ethnicities.

#3 *rs337711*: A large-scale genome-wide association study identified a correlation between *rs337711* and AF with a significant statistical P-value (Christophersen et al., 2017). However, another experiment based on vitro electrophysiology analysis and animal models failed to capture the association of this SNP with any atrial or ventricular changes in *KCNN2* mRNA expression (Bentzen et al., 2020). These contrasting results may lead to further exploration of the molecular mechanism of this SNP.

#4 *rs11264280*: Wang et al. (2018) conducted a clinical case study in the Chinese Han population but did not identify any significant correlation between *rs11264280* and AF. However, a later Mendelian randomization study conducted by Pan et al. (2020) indicates a notable correlation between this SNP and AF at a P-value of 3.07×10^{-79}, which is far smaller than the universal conspicuous level. Again, the conflicting results may suggest that this SNP performs differently for different ethnicities.

#5 *HP:* Eryd et al. (2011) conducted a cohort study to identify the association of haptoglobin (HP) level with AF, resulting in an insignificant level of correlation.

#6 *PKP2*: Several studies mention a potential association between gene *PKP2* and AF. Bourfiss et al. (2016) discovered that mutation in desmosomal *PKP2* could result in a significantly smaller atrial size in AF patients, which suggests a different arrhythmogenic mechanism of AF. Alhassani et al. (2018) reported a family case with large pathogenic *PKP2* deletion, resulting in cardiac arrhythmias, including persistent lone AF. These researchers also claim that AF occurring as genetic ventricular cardiomyopathy could be a secondary phenotype of a common underlying genetic variant.

#9 *rs3907*: The associations between AF and *rs3907* have barely been investigated at the current stage. According to the extremely

Table 6
Predicted disease-gene associations and evidence.

Rank	Gene/ Variant	DisGeNET *	P/ N §	Evidence¤	Evidence level†
#1	Gene | *BGLAP*	–	+	*BGLAP's* encoded protein's activity can be reduced by Warfarin, a treatment for AF (Sato et al., 2010; Yamagishi, 2019)	B
#2	SNP | *rs4762*	–	+	Identified as not significantly associated with AF by Zhao et al. (2015); Kuken et al. (2020) find the opposite.	B
#3	SNP | *rs337711*	0.700	+	Bentzen et al. (2020) and Christophersen et al. (2017) find no significant association; Wang et al. (2018) find the opposite.	A
#4	SNP | *rs11264280*	0.700	+	Wang et al. (2018) find no significant association; Pan et al. (2020) find the opposite.	A
#5	Gene | *HP*	–	+	Eryd et al. (2011) find no significant association between *HP's* protein product and the incidence of AF.	C
#6	Gene | *PKP2*	0.400	+	Yeung et al. (2019) find patients with mutations in desmosomal PKP2 have smaller atria. Alhassani et al. (2018) find large PKP2 deletion is related to lone AF.	A
#7	Gene | *DUOX2*	–	–	No clear findings	C
#8	Gene | *OLR1*	0.010	+	The activity of OCR1 increased by atrial modeling during AF (Bukowska et al., 2008)	A
#9	SNP | *rs3097*	–	–	Zhao et al. (2015) find no significant association.	C
#10	Gene | *MGP*	–	+	Warfarin treatment inhibits this gene expression	B
#11	Gene | *S100A6*	–	+	*S100A6*'s encoded protein combined with *RTEN* is a potential biomarker of AF (Doulamis et al., 2019)	B
#12	Gene | *IL3*	–	–	No clear findings	C
#13	Gene | *IL20*	–	–	No clear findings	C
#14	Gene | *TFF3*	–	+	*TFF3*'s encoded protein combined with *P3NP* is a potential biomarker of AF (Doulamis et al., 2019)	B
#15	Gene | *NOX4*	0.020	+	*NOX4* may be a gene therapy for ibrutinib-induced AF (Chen et al., 2019; Yang et al., 2020)	A

* The association score with AF retrieved from DisGeNET.
§ The overall evaluation of evidence identified from the literature, "+" indicates positive and "-" indicates negative.
¤ Sourced from LitVar, SNPedia &/or Literature Retrieval.
† A: positive evidence found both in DisGeNET and the literature; B: positive evidence found only in literature; C: neither the literature nor DisGeNET provide evidence of an association.
Sourced from LitVar, SNPedia &/or Literature Retrieval.

limited existing evidence (Zhao et al., 2015), this SNP is not significantly correlated with AF.
#11 *S100A6* and #14 *TFF-3*: A few studies have been conducted on the potential associations between these two genes and AF. From the limited evidence, it is possible that the genetic combinations of *TFF-3* & *P3NP* and *S100A6* & *RETN* may be biomarkers for AF (Doulamis et al., 2019)
#15 *NOX4*: Mounting evidence is revealing an association between *NOX4* and AF. Chen et al. (2019) evaluated the mediation of *CD44/NOX4* signals in atrial tachycardia-induced oxidative stress and Ca2+-handling abnormalities, providing a possible explanation for the onset/progression of AF. Yang et al. (2020) identified elevated expressions of *NOX4* in an ibrutinib-induced AF mice group, and proposed inhibiting *NOX* as a potential novel AF therapy for ibrutinib-induced AF.

We could not find supporting evidence for the other four predictions. However, in conversations with several domain experts, we were advised that the field is in a relatively early stage of research progress, and more time is needed to examine such associations. We believe these empirical insights can provide a clear direction for new, near-term research undertakings.

5. Discussion and conclusions

Taking individual genetic variability into account is a frontier trend in modern medical research. Awareness of a disease's genetic bases can contribute much to better risk assessment, diagnostics, and therapeutic treatment strategies. Our framework exploits the literature to capture known associations between biomedical entities and diseases and, further, combines the results with network analytics and link prediction to identify both core and potentially emerging genetic factors. The results of a comprehensive case study indicate our strategy is a very promising solution for genetic factor analysis and prediction.

The main contributions of our research include: 1) a cohesive methodology based on a combination of centrality and intersection ratio measurements for identifying the diseases, chemicals, genes, and genetic variants core to disease; 2) a semantic similarity-enhanced link prediction algorithm for generating more accurate predictions of the possible associations between genes and diseases; 3) an adaptable and transferable framework for general use in genetic factor analysis and prediction.

Our empirical study focused on atrial fibrillation. The results of the case analysis presented some common entities associated with AF. However, they also revealed some controversial findings, such as the association between AF and gastroesophageal reflux, Omega-3 fatty acids, and caffeine. Therefore, from one perspective, the framework can be seen as a tool for generating a data-driven, bird's eye view of cardiovascular research. From another perspective, it is a decision support system that generates insights into prior research that may need to be re-examined or pointers toward future research that is likely to prove fruitful.

There are also several limitations of this paper that may require investigation in future studies. The first comes from a technical standpoint. We emphasized the need to identify strong associations by adopting co-occurrence analysis. However, this process inevitably retrieves negative associations along with the positives and because it does not recognize the causalities between entities. Embeddings of SPO triples could be a challenging but significant task in this area, and we have conducted certain pilot studies on this trail (Zhang et al., 2020c) by incorporating word embeddings with SPO triples and it is foreseeable to enhance its capabilities in identifying the causalities of entities. Second, due to the lack of available rules for entity extraction, we dealt with disease entities at the granularity defined in MeSH and did not further classify disease entities into other subtypes. Yet, in the context of biomedical entity inference, the molecular mechanisms that underpin the different types of atrial fibrillation are quite diverse. A rule-based

filter could overcome this problem, which is also on our agenda. From a biomedical standpoint, there is also an issue with selecting the core entities by means of inference. Guilt-by-association is a prevalent hypothesis for establishing genetic associations for diseases. Thus, we did instinctively filter out the neighbors of the core entities to narrow down the model input, but it does not contribute significantly to our recall performance. Despite this limitation, however, we do believe that emphasizing the core genetic factors for the purposes of link prediction is still a promising approach for improving performance. A solution to this problem is going to require deeper on a range of alternative approaches.

Acknowledgements

Pilot studies of part of this work have been submitted to the 2020/2021 Portland International Center for Management of Engineering and Technology (PICMET 2020/2021) and the workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE 2020) at the 2020 ACM/IEEE Joint Conference on Digital Libraries (JCDL 2020).

This work is supported by the Australian Research Council under Discovery Early Career Researcher Award DE190100994.

References

Abdelfattah, R., Kamran, H., Lazar, J., Kassotis, J., 2018. Does caffeine consumption increase the risk of new-onset atrial fibrillation? Cardiology 140, 106–114.

Abdelhamid, A.S., Brown, T.J., Brainard, J.S., Biswas, P., Thorpe, G.C., Moore, H.J., et al., 2020. Omega-3 fatty acids for the primary and secondary prevention of cardiovascular disease. Cochrane Database Systemat. Rev. (3) https://doi.org/10.1002/14651858.CD003177.pub5.

Adamic, L.A., Wilkinson, D., Huberman, B.A., Adar, E., 2002. A literature based method for identifying gene-disease connections. In: Paper presented at the Proceedings. IEEE Computer Society Bioinformatics Conference.

Al-Aamri, A., Taha, K., Al-Hammadi, Y., Maalouf, M., Homouz, D., 2019. Analyzing a co-occurrence gene-interaction network to identify disease-gene association. BMC Bioinformatics 20 (1), 70.

Alhassani, S., Deif, B., Conacher, S., Cunningham, K.S., Roberts, J.D., 2018. A large familial pathogenic Plakophilin-2 gene (PKP2) deletion manifesting with sudden cardiac death and lone atrial fibrillation: evidence for alternating atrial and ventricular phenotypes. HeartRhythm Case Rep. 4 (10), 486–489.

Anastasiadou, E., Jacob, L.S., Slack, F.J., 2018. Non-coding RNA networks in cancer. Nature Rev. Cancer 18 (1), 5.

Arias, T.D., Jorge, L., Barrantes, R., 1991. Uses and misuses of definitions of genetic polymorphism. A perspective from population pharmacogenetics. Br. J. Clin. Pharmacol. 31 (1), 117.

Barabási, A.-.L., Gulbahce, N., Loscalzo, J., 2011. Network medicine: a network-based approach to human disease. Nature Rev. Genetic. 12 (1), 56–68.

Bentzen, B.H., Bomholtz, S.H., Simó-Vicens, R., Folkersen, L., Abildgaard, L., Speerschneider, T., et al., 2020. Mechanisms of Action of the KCa2-Negative Modulator AP30663, a Novel Compound in Development for Treatment of Atrial Fibrillation in Man. Front. Pharmacol. 11, 610.

Bourfiss, M., Te Riele, A.S., Mast, T.P., Cramer, M.J., Van Der Heijden, J.F., Van Veen, T. A., et al., 2016. Influence of genotype on structural atrial abnormalities and atrial fibrillation or flutter in arrhythmogenic right ventricular dysplasia/cardiomyopathy. J. Cardiovasc. Electrophysiol. 27 (12), 1420–1428.

Bukowska, A., Schild, L., Keilhoff, G., Hirte, D., Neumann, M., Gardemann, A., et al., 2008. Mitochondrial dysfunction and redox signaling in atrial tachyarrhythmia. Exp. Biol. Med. 233 (5), 558–574.

Bush, W.S., Moore, J.H., 2012. Genome-wide association studies. PLoS Comput. Biol. 8 (12).

Cariaso, M., Lennon, G., 2012. SNPedia: a wiki supporting personal genome annotation, interpretation and analysis. Nucleic Acids Res. 40 (D1), D1308–D1312.

Chen, W.-.J., Chang, S.-.H., Chan, Y.-.H., Lee, J.-.L., Lai, Y.-.J., Chang, G.-.J., et al., 2019. Tachycardia-induced CD44/NOX4 signaling is involved in the development of atrial remodeling. J. Mol. Cell. Cardiol. 135, 67–78.

Christophersen, I.E., Rienstra, M., Roselli, C., Yin, X., Geelhoed, B., Barnard, J., et al., 2017. Large-scale analyses of common and rare variants identify 12 new loci associated with atrial fibrillation. Nat. Genet. 49 (6), 946–952.

Clemente-Casares, X., Blanco, J., Ambalavanan, P., Yamanouchi, J., Singha, S., Fandos, C., et al., 2016. Expanding antigen-specific regulatory networks to treat autoimmunity. Nature 530 (7591), 434–440.

Cohen, A.M., Hersh, W.R., Dubay, C., Spackman, K., 2005. Using co-occurrence network structure to extract synonymous gene and protein names from MEDLINE abstracts. BMC Bioinformatics 6 (1), 103.

Cookson, W., Liang, L., Abecasis, G., Moffatt, M., Lathrop, M., 2009. Mapping complex disease traits with global gene expression. Nature Rev. Genetics 10 (3), 184–194.

Coulet, A., Shah, N.H., Garten, Y., Musen, M., Altman, R.B., 2010. Using text to build semantic networks for pharmacogenomics. J. Biomed. Inform. 43 (6), 1009–1019.

Curatolo, P.W., Robertson, D., 1983. The health consequences of caffeine. Ann. Intern. Med. 98 (5_Part_1), 641–653.

Doulamis, I.P., Samanidis, G., Tzani, A., Antoranz, A., Gkogkos, A., Konstantopoulos, P., et al., 2019. Proteomic profile of patients with atrial fibrillation undergoing cardiac surgery. Interact Cardiovasc. Thorac. Surg. 28 (1), 94–101.

Düzen, I.V., Yavuz, F., Vuruskan, E., Saracoglu, E., Poyraz, F., Göksülük, H., et al., 2017. Leukocyte TRP channel gene expressions in patients with non-valvular atrial fibrillation. Sci. Rep. 7 (1), 1–7.

Ellinor, P.T., Lunetta, K.L., Glazer, N.L., Pfeufer, A., Alonso, A., Chung, M.K., et al., 2010. Common variants in KCNN3 are associated with lone atrial fibrillation. Nat. Genet. 42 (3), 240–244.

Eryd, S.A., Smith, J.G., Melander, O., Hedblad, B., Engström, G., 2011. Inflammation-sensitive proteins and risk of atrial fibrillation: a population-based cohort study. Eur. J. Epidemiol. 26 (6), 449.

Feghaly, J., Zakka, P., London, B., MacRae, C.A., Refaat, M.M., 2018. Genetics of atrial fibrillation. J. Am. Heart Assoc. 7 (20), e009884.

Freeman, L.C., Roeder, D., Mulholland, R.R., 1979. Centrality in social networks: II. Experimental results. Soc Netw. 2 (2), 119–141.

Ganegoda, G.U., Wang, J., Wu, F.-.X., Li, M., 2014. Prediction of disease genes using tissue-specified gene-gene network. BMC Syst. Biol. 8 (S3), S3.

Garten, Y., Tatonetti, N.P., Altman, R.B., 2010. Improving the prediction of pharmacogenes using text-derived drug-gene relationships. In: Biocomputing 2010. World Scientific, pp. 305–314.

Goldstein, D.B., 2009. Common genetic variation and human traits. New Engl. J. Med. 360 (17), 1696.

Habibi, M., Weber, L., Neves, M., Wiegandt, D.L., Leser, U., 2017. Deep learning with word embeddings improves biomedical named entity recognition. Bioinformatics 33 (14), i37–i48.

Heo, G.E., Xie, Q., Song, M., Lee, J.-.H., 2019. Combining entity co-occurrence with specialized word embeddings to measure entity relation in Alzheimer's disease. BMC Med. Inform. Decis. Mak. 19 (5), 240.

Huang, T.C., Lo, L.W., Yamada, S., Chou, Y.H., Lin, W.L., Chang, S.L., et al., 2019. Gastroesophageal reflux disease and atrial fibrillation: insight from autonomic cardiogastric neural interaction. J. Cardiovasc. Electrophysiol. 30 (11), 2262–2270.

Jenssen, T.-.K., Lægreid, A., Komorowski, J., Hovig, E., 2001. A literature network of human genes for high-throughput analysis of gene expression. Nat. Genet. 28 (1), 21–28.

Kim, J., Kim, J.-j., Lee, H., 2017. An analysis of disease-gene relationship from Medline abstracts by DigSee. Sci. Rep. 7 (1), 1–13.

Kuken, B., Yang, Y., Liu, Z., He, P., Wulasihan, M., 2020. Relationship between M235T and T174M polymorphisms in angiotensin gene and atrial fibrillation in Uyghur and Han populations of Xinjiang, China. Int. J. Clin. Exp. Pathol. 13 (8), 2065.

Landrum, M.J., Lee, J.M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., et al., 2016. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. 44 (D1), D862–D868.

Lawson, C.E., Wu, S., Bhattacharjee, A.S., Hamilton, J.J., McMahon, K.D., Goel, R., et al., 2017. Metabolic network analysis reveals microbial community interactions in anammox granules. Nat. Commun. 8 (1), 1–12.

Lei, C., Ruan, J., 2013. A novel link prediction algorithm for reconstructing protein–protein interaction networks by topological similarity. Bioinformatics 29 (3), 355–364.

Li, Y.-y., Zhou, C.-w., Xu, J., Qian, Y., Wang, B., 2012. CYP11B2 T-344C gene polymorphism and atrial fibrillation: a meta-analysis of 2,758 subjects. PLoS ONE 7 (11), e50910.

Liben-Nowell, D., Kleinberg, J., 2007. The link-prediction problem for social networks. J. Am. Soc. Inf. Sci. Technol. 58 (7), 1019–1031.

Lovász, L., 1993. Random walks on graphs: a survey. Combinatorics, Paul erdos is eighty 2 (1), 1–46.

Lü, L., Jin, C.-.H., Zhou, T., 2009. Similarity index based on local paths for link prediction of complex networks. Phys. Rev. E 80 (4), 046122.

Lü, L., Zhou, T., 2010. Link prediction in weighted networks: the role of weak ties. EPL (Europhysics Letters) 89 (1), 18001.

Mallory, E.K., Zhang, C., Ré, C., Altman, R.B., 2016. Large-scale extraction of gene interactions from full-text literature using DeepDive. Bioinformatics 32 (1), 106–113.

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Newburger, D.E., Bulyk, M.L., 2009. UniPROBE: an online database of protein binding microarray data on protein–DNA interactions. Nucleic Acids Res. 37 (suppl_1), D77–D82.

Niwattanakul, S., Singthongchai, J., Naenudorn, E., Wanapu, S., 2013. Using of Jaccard coefficient for keywords similarity. In: Paper presented at the Proceedings of the international multiconference of engineers and computer scientists.

Olesen, M.S., Bentzen, B.H., Nielsen, J.B., Steffensen, A.B., David, J.-.P., Jabbari, J., et al., 2012. Mutations in the potassium channel subunit KCNE1 are associated with early-onset familial atrial fibrillation. BMC Med. Genet. 13 (1), 1–9.

Opap, K., Mulder, N., 2017. Recent advances in predicting gene-disease associations. F1000Res 6. https://doi.org/10.12688/f1000research.10788.1, 578-578.

Ott, J., 1999. Analysis of Human Genetic Linkage. JHU Press.

Özgür, A., Vu, T., Erkan, G., Radev, D.R., 2008. Identifying gene-disease associations using centrality on a literature mined gene-interaction network. Bioinformatics 24 (13), i277–i285.

Pan, Y., Wang, Y., Wang, Y., 2020. Investigation of causal effect of atrial fibrillation on Alzheimer disease: a mendelian randomization study. J. Am. Heart Assoc 9 (2), e014889.
Piñero, J., Bravo, À., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., et al., 2016. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res. gkw943.
Pletscher-Frankild, S., Pallejà, A., Tsafou, K., Binder, J.X., Jensen, L.J., 2015. DISEASES: text mining and data integration of disease–gene associations. Methods 74, 83–89.
Rong, X., 2014. word2vec parameter learning explained. arXiv preprint arXiv: 1411.2738.
Roselli, C., Rienstra, M., Ellinor, P.T., 2020. Genetics of atrial fibrillation in 2020: GWAS, genome sequencing, polygenic risk, and beyond. Circ. Res. 127 (1), 21–33.
Sato, Y., Honda, Y., Jun, I., 2010. Long-term oral anticoagulation therapy and the risk of hip fracture in patients with previous hemispheric infarction and nonrheumatic atrial fibrillation. Cerebrovascul. Diseases 29 (1), 73–78.
Schellenberger, J., Park, J.O., Conrad, T.M., Palsson, B.Ø., 2010. BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. BMC Bioinformatics 11 (1), 213.
Schlicker, A., Lengauer, T., Albrecht, M., 2010. Improving disease gene prioritization using the semantic similarity of Gene Ontology terms. Bioinformatics 26 (18), i561–i567.
Sheikh, O., Vande Hei, A.G., Battisha, A., Hammad, T., Pham, S., Chilton, R., 2019. Cardiovascular, electrophysiologic, and hematologic effects of omega-3 fatty acids beyond reducing hypertriglyceridemia: as it pertains to the recently published REDUCE-IT trial. Cardiovasc. Diabetol. 18 (1), 84. https://doi.org/10.1186/s12933-019-0887-0.
Sinner, M.F., Pfeufer, A., Akyol, M., Beckmann, B.-.M., Hinterseer, M., Wacker, A., et al., 2008. The non-synonymous coding IKr-channel variant KCNH2-K897T is associated with atrial fibrillation: results from a systematic candidate gene-based analysis of KCNH2 (HERG). Eur. Heart J. 29 (7), 907–914.
Stapley, B.J., Benoit, G., 2000. Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts. Pac. Symp. Biocomput 529–540. https://doi.org/10.1142/9789814447331_0050.
Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., et al., 2019. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic. Acids Res. 47 (D1), D607–D613.
Tong, H., Faloutsos, C., Pan, J.-.Y., 2008. Random walk with restart: fast solutions and applications. Knowl. Inf. Syst. 14 (3), 327–346.
Valdeolivas, A., Tichit, L., Navarro, C., Perrin, S., Odelin, G., Levy, N., et al., 2019. Random walk with restart on multiplex and heterogeneous biological networks. Bioinformatics 35 (3), 497–505.
van Dam, S., Vosa, U., van der Graaf, A., Franke, L., de Magalhaes, J.P., 2018. Gene co-expression analysis for functional classification and gene–disease predictions. Brief. Bioinformatics 19 (4), 575–592.
Wang, J., Liu, Q., Yuan, S., Xie, W., Liu, Y., Xiang, Y., et al., 2017. Genetic predisposition to lung cancer: comprehensive literature integration, meta-analysis, and multiple evidence assessment of candidate-gene association studies. Sci. Rep. 7 (1), 1–13.
Wang, X., Nie, Y., Ning, S., Shi, Y., Zhao, Y., Niu, S., et al., 2018. Rs17042171 at chromosome 4q25 is associated with atrial fibrillation in the Chinese Han population from the central plains. J. Central South Univer. Med. Sci. 43 (6), 594.
Wei, C.-.H., Allot, A., Leaman, R., Lu, Z., 2019. PubTator central: automated concept annotation for biomedical full text articles. Nucleic. Acids Res. 47 (W1), W587–W593.
Wei, C.-.H., Kao, H.-.Y., Lu, Z., 2013. PubTator: a web-based text mining tool for assisting biocuration. Nucleic. Acids Res. 41 (W1), W518–W522.
Xie, W.-.H., Chang, C., Xu, Y.-.J., Li, R.-.G., Qu, X.-.K., Fang, W.-.Y., et al., 2013. Prevalence and spectrum of Nkx2. 5 mutations associated with idiopathic atrial fibrillation. Clinics 68 (6), 777–784.
Yamagishi, S.-i, 2019. Concerns about clinical efficacy and safety of Warfarin in diabetic patients with atrial fibrillation. Cardiovasc. Diabetol. 18 (1), 12.
Yang, X., An, N., Zhong, C., Guan, M., Jiang, Y., Li, X., et al., 2020. Enhanced cardiomyocyte reactive oxygen species signaling promotes ibrutinib-induced atrial fibrillation. Redox Biol. 30, 101432.
Yeung, C., Enriquez, A., Suarez-Fuster, L., Baranchuk, A., 2019. Atrial fibrillation in patients with inherited cardiomyopathies. Ep Europace 21 (1), 22–32.
Yuan, Y., Xu, H., Wang, B., 2014. An improved NSGA-III procedure for evolutionary many-objective optimization. In: Paper presented at the Proceedings of the 2014 Annual Conference on Genetic and Evolutionary Computation.
Zhang, Y., Lu, J., Liu, F., Liu, Q., Porter, A., Chen, H., et al., 2018a. Does deep learning help topic extraction? A kernel k-means clustering method with word embedding. J. Informetr. 12 (4), 1099–1117.
Zhang, Y., Porter, A.L., Cunningham, S., Chiavetta, D., Newman, N., 2020a. Parallel or Intersecting Lines? Intelligent Bibliometrics for Investigating the Involvement of Data Science in Policy Analysis. IEEE Trans. Eng. Manag.
Zhang, Y., Qian, Y., Huang, Y., Guo, Y., Zhang, G., Lu, J., 2017. An entropy-based indicator system for measuring the potential of patents in technological innovation: rejecting moderation. Scientometrics 111 (3), 1925–1946.
Zhang, Y., Wang, X., Zhang, G., Lu, J., 2018b. Predicting the dynamics of scientific activities: a diffusion-based network analytic methodology. Proc. Associat. Inf. Sci. Technol. 55 (1), 598–607. https://doi.org/10.1002/pra2.2018.14505501065.
Zhang, Y., Wu, M., Hu, Z., Ward, R., Zhang, X., Porter, A., 2020b. Profiling and predicting the problem-solving patterns in China's research systems: a methodology of intelligent bibliometrics and empirical insights. Q. Sci. Stud. accepted on 18/09/2020.
Zhang, Y., Wu, M., Zhu, Y., Huang, L., Lu, J., 2020c. Characterizing the potential of emerging generic technologies: a methodology based on bi-layer network analytics. J. Informetr. under review.
Zhang, Y., Zhang, G., Chen, H., Porter, A.L., Zhu, D., Lu, J., 2016. Topic analysis and forecasting for science, technology and innovation: methodology with a case study focusing on big data research. Technol. Forecast. Soc. Change 105, 179–191. https://doi.org/10.1016/j.techfore.2016.01.015.
Zhao, L.-q., Wen, Z.-j., Wei, Y., Xu, J., Chen, Z., Qi, B.-z., et al., 2015. Polymorphisms of renin-angiotensin-aldosterone system gene in Chinese Han patients with nonfamilial atrial fibrillation. PLoS ONE 10 (2), e0117489.
Zhou, T., Lü, L., Zhang, Y.-.C., 2009. Predicting missing links via local information. Eur. Phys. J. B 71 (4), 623–630.

Mengjia Wu: Mengjia Wu is working toward Ph.D. degree at the Australian Artificial Intelligence Institute, Faculty of Engineering and Information Technology, University of Technology Sydney, Australia. He has published papers in several conferences including PICMET, ISKE and ISSI. His-research interests include bibliometrics, text mining, network analytics and information management.

Yi Zhang: Yi Zhang received his dual PhD degrees in Management Science & Engineering from Beijing Institute of Technology 2016 and in Software Engineering from University of Technology Sydney (UTS) 2017. He is a Lecturer in the UTS Australian Artificial Intelligence Institute. His-research interests include bibliometrics, text analytics, and innovation & technology management. Dr. Zhang serves diverse roles (e.g., Associate Editor, Editorial Board Member, and Managing Guest Editor) in IEEE Trans and other international journals. He is a member of the Advisory Board for ICSR, and member of ASIS&T, ISSI, and IEEE. He was the recipient of the 2019 Discover Early Career Researcher Award granted by the Australian Research Council.

Guangquan Zhang: Guangquan Zhang is an Associate Professor and Director of the Decision Systems and e-Service Intelligent (DeSI) Research Laboratory at the University of Technology Sydney Australia. He received the Ph.D. degree in applied mathematics from Curtin University of Technology, Australia, in 2001. His-research interests include fuzzy machine learning, fuzzy optimization, and machine learning. He has authored five monographs, five textbooks, and 460 papers including 220 refereed international journal papers. Dr. Zhang has won seven Australian Research Council (ARC) Discovery Projects grants and many other research grants. He was awarded an ARC QEII fellowship in 2005. He has served as a member of the editorial boards of several international journals, as a guest editor of eight special issues for IEEE transactions and other international journals, and co-chaired several international conferences and workshops in the area of fuzzy decision-making and knowledge engineering.

Jie Lu: Jie Lu is a Distinguished Professor and the Director of Australian Artificial Intelligence Institute (AAII) at the University of Technology Sydney (UTS), Australia. She is also an IFSA Fellow and Australian Laureate Fellow. She received a PhD degree from Curtin University, Australia, in 2000. Her main research expertise is in fuzzy transfer learning, concept drift, decision support systems and recommender systems. She has been awarded 10 Australian Research Council (ARC) discovery grants and led 15 industry projects. She has supervised 40 PhD students to completion. She serves as Editor-In-Chief for Knowledge-Based Systems (Elsevier) and Editor-In-Chief for International Journal on Computational Intelligence Systems (Atlantis). She has delivered 27 keynote speeches at IEEE and other international conferences and chaired 15 international conferences. She has received the UTS Medal for Research and Teaching Integration (2010), the UTS Medal for research excellence (2019), the IEEE Transactions on Fuzzy Systems Outstanding Paper Award (2019), the Computer Journal Wilkes Award (2018), and the Australian Most Innovative Engineer Award (2019).